NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Recolha de contratos de despesa pública e segmentação dos perfis de despesa a nível Municipal

Filipe Manuel Leitão Gonçalves Freire

Dissertação como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AGRADECIMENTOS

Com os meus reconhecidos agradecimentos ao meu orientador, Professor Doutor Flávio L. Pinheiro pelo inestimável apoio prestado na orientação e supervisão desta proposta de tese.
Ainda quero agradecer à minha família pelo suporte e compreensão prestados ao longo deste trabalho.

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

RECOLHA DE CONTRATOS DE DESPESA PÚBLICA E SEGMENTAÇÃO DOS PERFIS DE DESPESA A NÍVEL MUNICIPAL

por

Filipe Manuel Leitão Gonçalves Freire

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação Especialização em Gestão do Conhecimento e Business Intelligence.

Orientador: Professor Doutor Flávio Luís Portas Pinheiro

24 de Novembro de 2019

RESUMO

Devido à necessidade de analisar como são investidos os capitais públicos nos municípios Portugueses nos diversos tipos de contratos de aquisição de bens e serviços, torna-se fundamental criar ferramentas que permitam a compreensão destes investimentos. É desejável perceber como oscilam estes investimentos em função da dimensão da população.

Neste projeto, o objetivo é recolher dados disponibilizados na web sobre contratos e criar uma segmentação para os diversos tipos de despesa pública, que permita detetar eventuais desvios anómalos na relação entre despesa pública municipal e dimensão populacional.

Para este efeito, foi desenvolvido um *web crawler* com recurso à linguagem de programação *Python* que permitiu extrair de forma automática os contratos públicos do site *http://www.base.gov.pt/*. Foram analisados os dados recolhidos tendo sido detetada uma relação do tipo *log-log* entre população e despesa pública. Posteriormente foi feita uma análise de segmentação com base nos resíduos da relação anteriormente mencionada com recurso a técnicas de *DataMining.* Foram usados diversos algoritmos de *Clustering*, em particular, o *K-Medoids,* do qual foram gerados dois grupos distintos de tipos de despesa.

PALAVRAS-CHAVE

Contratos Públicos; Web Scraping; Clustering; Data Mining; Python

ABSTRACT

Due to the need to analyze how public capital is invested in Portuguese municipalities in the various types of contracts for the acquisition of goods and services, it is essential to create tools that allow the understanding of these investments. It is desirable to understand how these investments oscillate according to the size of the population.

In this project, the objective is to collect data available on the web about contracts and to create a segmentation for the various types of public expenditure, allowing to detect any anomalous deviations in the relationship between municipal public expenditure and population size.

For this purpose, a web crawler was developed using the Python programming language that allowed to automatically extract public contracts from the site http://www.base.gov.pt/. The data collected were analyzed and a *log-log* relationship between population and public expenditure was detected. Subsequently, a segmentation analysis based on the residues of the referred relationship was performed using DataMining techniques. Several *Clustering* algorithms were used, in particular *K-Medoids*, from which two distinct groups of expense types were generated.

KEYWORDS

Public Procurements; Web Scraping; Clustering; Data Mining; Python

Índice

Índice de Tabelas

Índice de Figuras

Índice de Anexos

1.Introdução

A despesa pública pretende colmatar as necessidades coletivas duma comunidade, que podem ser de diferentes tipos tais como infraestruturas físicas, serviços de saúde, educativos, sociais ou de segurança. Para poder efetuar despesa pública os governos nacionais, regionais ou locais necessitam de fundos (capital), que normalmente advém de taxas cobradas ou de empréstimos, advindo daqui uma tensão entre os Governos e os contribuintes. A despesa pública embora tenha como objetivo teórico beneficiar a comunidade, na prática, pode ser boa ou má consoante a relação custo/benefício.

O economista Keynes (Keynes, 1936) considerou que a despesa pública pode ser um fator fundamental para dinamizar a economia quando esta se encontra num período de recessão. Por outro lado, as correntes neo-liberais (Friedman & Friedman, 1962) têm desconfiança da intervenção do estado, pois consideram que o estado não está livre de enviesamentos políticos e advogam que a alocação de recursos, determinada pela oferta e procura, é menos sujeita a erros do que quando determinada por políticos ou burocratas.

A despesa pública é um tema recorrente quando se fala da gestão económica de um país ou região, pois daí decorre em grande parte a qualidade dos serviços que são postos à disposição da população. A despesa pública carece de ser analisada de modo a perceber a alocação da despesa aos diferentes domínios da atividade económica, tendo em conta diversos fatores, tais como densidade populacional, localização geográfica, riqueza da região, nível de desenvolvimento económico, nível de desemprego ou outras questões sociais. Para além da despesa central do Estado, existe a despesa a nível municipal que é pertinente analisar para se ter uma ideia de como funciona o investimento a nível local, quais os tipos de despesa que tipicamente são efetuados e de que forma essa despesa é excessiva ou insuficiente.

Uma parte da despesa pública realizada é a contratação pública, que em Portugal representa cerca de 19,5% da despesa pública total e 10% do PIB[1]. A contratação pública tem origem nas diversas instituições da Administração Pública tais como o Estado, as Autarquias locais, os Institutos

[1] http://www.concorrencia.pt/vPT/Noticias_Eventos/Intervencoes_publicas/Documents/Contrata%C3%A7%C3%A3o%20P%C3%BAblica%20e%20Concorr%C3%AAncia%20-%20a%20import%C3%A2ncia%20do%20di%C3%A1logo%20interinstitucional.pptx

públicos, entre outros. Os procedimentos de contratação pública são vários: concurso público, ajuste direto, consulta prévia, concurso limitado por prévia qualificação, negociação, diálogo concorrencial e parceria para a inovação. Em Portugal, a contratação pública realizada pelas diversas entidades está sujeita às regras definidas pelo Código dos Contratos Públicos (CCP), deve ter em conta a oferta do mercado, podendo ser abrangidos contratos tais como: empreitadas de obras, concessão de obras públicas, concessão de serviços públicos, locação ou aquisição de bens móveis e aquisição de serviços.

A contratação pública obedece a legislação que tem que ser seguida pelas diversas entidades do Estado. Neste âmbito a Comunidade Europeia decidiu criar um sistema de classificação da contratação pública de forma a tipificar essa despesa, permitindo uma maior transparência e comparabilidade em relação aos investimentos. O facto de em Portugal existirem trezentos e oito municípios com elevada variedade de número de habitantes e de todos eles contribuírem para a contratação pública motiva fazer um estudo utilizando ferramentas estatísticas e de Data Mining procurando discernir quais os perfis de despesa em que se verificam mais desvios face à tendência considerando a relação investimento versus população. A possibilidade de fazer este estudo depende da disponibilização dos dados dos contratos públicos efetuados pelas diversas entidades. O Estado Português disponibiliza os dados dos contratos realizados pela administração pública num Portal online, onde uma complexa e detalhada informação referente aos contratos podem ser consultadas publicamente. Este portal foi concebido para um acesso interativo via Browser Web, no entanto usando tecnologia adequada de Web Scraping desenhou-se um processo para conseguir a recolha massiva dos dados.

1.1.Enquadramento Teórico

Durante o século XX assistimos a um crescimento significativo da despesa pública a nível mundial. Várias teorias chamadas de Leis de Evolução da Despesa Pública, visam explicar este fenómeno como *lei de Wagner, dos efeitos de deslocação e da teoria da produtividade diferencial* (Almeida, 2001).

Em Portugal, o crescimento das despesas públicas ocorre sobretudo em dois momentos, o primeiro entre 1940 e 1970 que pode ser descrito como "lento" e o segundo a partir de 1970 onde houve uma aceleração significativa (**Figura 1**).

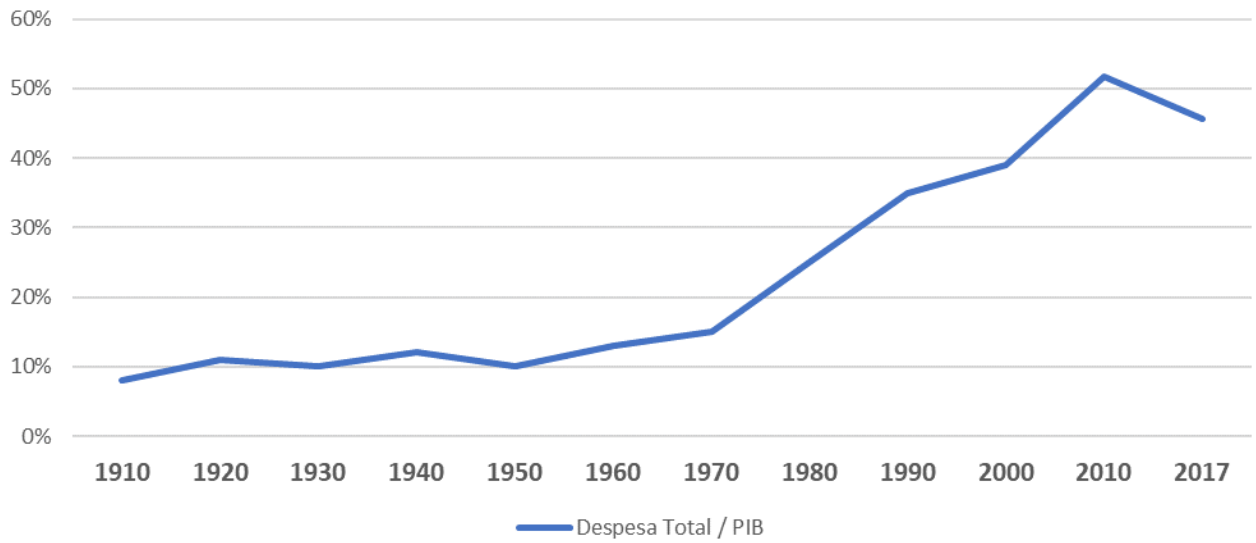


Figura 1 - % Despesa total pública sobre o PIB em Portugal (1990-2017)[2]

Ao contrário do resto da Europa, esta segunda fase do crescimento chega a Portugal com pelo menos uma década de atraso devido às condições económicas e sociais (Tanzi & Schuknecht, 2000). Mais recentemente em 2017, Portugal apresentou um valor de 45,7% da despesa pública sobre o PIB, valor alinhado com a UE que segundo a Eurostat pesa 45,8% tornando óbvio que a relação entre adjudicantes e adjudicatários deve ser o mais transparente e eficiente possível (Kuljanin & Klipstein, 2017).

Em 1993, a comunidade Europeia cria a primeira versão do atual Vocabulário Comum para os Contratos Públicos, denominados de *Community Procuremment Vocabulary* (CPV) para

[2] CGE; Santos (1984); Séries Longas INE e Banco de Portugal *in* Mateus (1998); (pordata, s.d.)

classificar os contratos públicos. Esta classificação permite que as diversas entidades quer adjudicantes quer adjudicatárias tenham referências normalizadas para caraterizar o objeto dos contratos, facilitando assim o desenvolvimento de estudos comparativos entre diferentes países e instituições. Esta classificação é oriunda dos diversos sistemas de classificação existentes na altura: *Central Product Classification* (CPC), *International Standard Industrial Classification* (ISIC) e *Classification of Products by Activity* (CPA). Estas classificações eram compostas por 6 dígitos e mais direcionadas para os fornecedores de serviços do que para os adjudicantes, continuado assim a haver uma clara insuficiência no que toca à agilidade no momento da contração do serviço ou compra de produtos.

Em 1994, embora a sigla CPV se tenha mantido, o nome é alterado para *Common Procuremment Vocabulary*. Entre 1994 e 2008 o CPV sofreu várias restruturações principalmente na adição e remoção de novas categorias. A última atualização feita na estrutura foi em 2008 onde o anexo I do Regulamento (CE) Nº 213/2008 atualiza o Vocabulário Comum para os Contratos Públicos (CPV) definindo códigos classificativos de 8 algarismos. Estes códigos definem uma estrutura em árvore. A estrutura do código de 8 algarismos divide-se numa parte principal formada por 5 algarismos e numa parte suplementar de 3 algarismos. A parte principal define os primeiros quatro níveis da seguinte forma:

- **Divisões** - dois primeiros algarismos;
- **Grupos** - três primeiros algarismos;
- **Classes** - quatro primeiros algarismos;
- **Categorias** - cinco primeiros algarismos.

No total existem **9454** códigos distribuídos da seguinte forma (**Figura 2**):

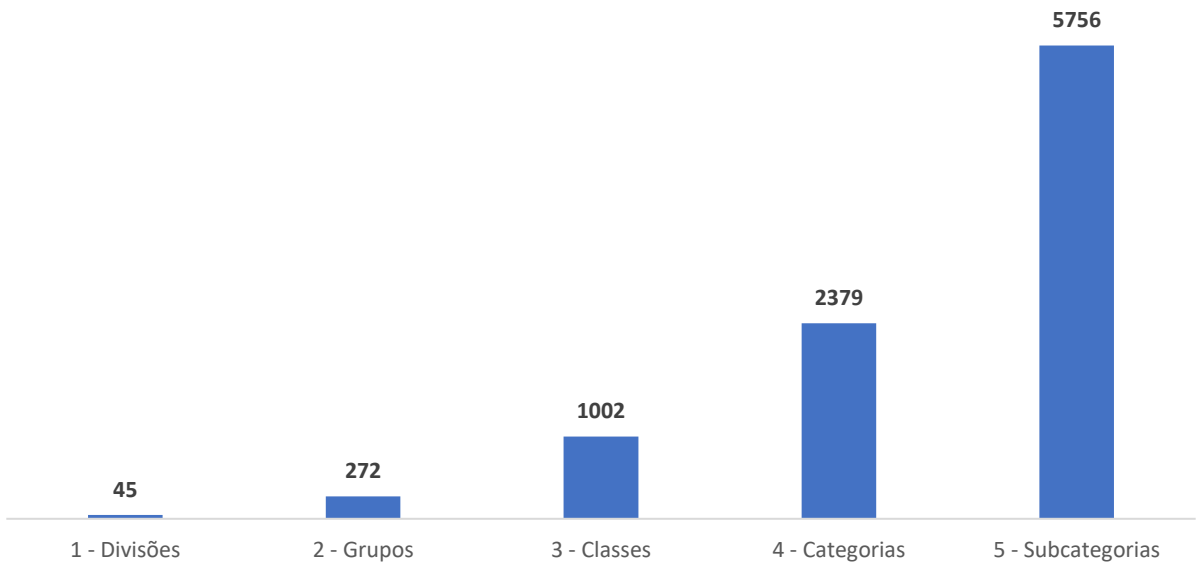


Figura 2 - Distribuição do número de códigos no CPV

Em adição ao Vocabulário Comum para os Contratos Públicos, foi criado um vocabulário secundário cujo o objetivo é poder detalhar mais a descrição do contrato. Segundo o relatório requerido pela Comissão Europeia em 2012 sobre o funcionamento dos códigos CPV, este vocabulário secundário tem uma taxa de utilização na ordem dos 1,5%. A estrutura de códigos, é composta por duas letras e dois algarismos formando um total de 903 códigos. Ainda segundo o relatório anteriormente mencionado, quando comparado com outros sistemas classificativos, o sistema CPV apesar de globalmente abranger mais serviços e categorias é menos detalhado para alguns níveis e apresenta falta de consistência na estrutura hierárquica assim como categorias com descrições não mutuamente disjuntas.

Em 2008, O Decreto-Lei n.º 18/2008 prevê a construção de um portal único com informação relativa aos contratos Públicos bem como a sua gestão e autenticação. Esta medida tem como objetivo aumentar a transparência num setor de elevada importância. Ainda refere segundo o Artigo 1º a criação de um portal na Internet dedicado à disponibilização e gestão de todos os contratos públicos bem como a disponibilização de informações estatísticas ou relevantes.

Em Portugal, o Portal BASE, é o portal com a informação sobre todos os contratos públicos Portugueses celebrados pelas diversas entidades da administração pública. No Portal BASE

encontra-se informação sobre procedimentos, sendo possível a consulta dos anúncios que estão em vigor. De forma a cumprir as diretrizes previstas pela Comissão Europeia compete ao Instituto dos Mercados Públicos, do Imobiliário e da Construção (IMPIC) a elaboração de relatórios estatísticos relativos aos contratos públicos. O IMPIC ainda é responsável por toda gestão e monotorização do portal BASE. A classificação dos contratos através de códigos supõe a atribuição do contrato a um determinado tipo de atividade económica. Esta classificação pode ser feita a um nível de detalhe mais ou menos elevado. Nesta análise consideram-se apenas os dois primeiros dígitos do código CPV que definem 45 tipos de atividade económica (Anexo 1).

Considerando a globalidade da despesa pública total efetuada em contratos públicos ao longo dos anos 2009-2017 verifica-se que a despesa não apresenta uma distribuição homogénea pelos diferentes códigos quer em investimento quer em número de contratos. Em particular verifica-se que o tipo de despesa **45 – Construção** apresenta um valor de investimento claramente mais elevado. Por outro lado, o tipo de despesa **33 – Equipamento médico**, medicamentos e produtos para cuidados pessoais apresenta um maior número de contratos realizados.

As **Figura 3** e **Figura 4** representam a distribuição de valor e despesa pela tipologia de CPV's (2 Dígitos).

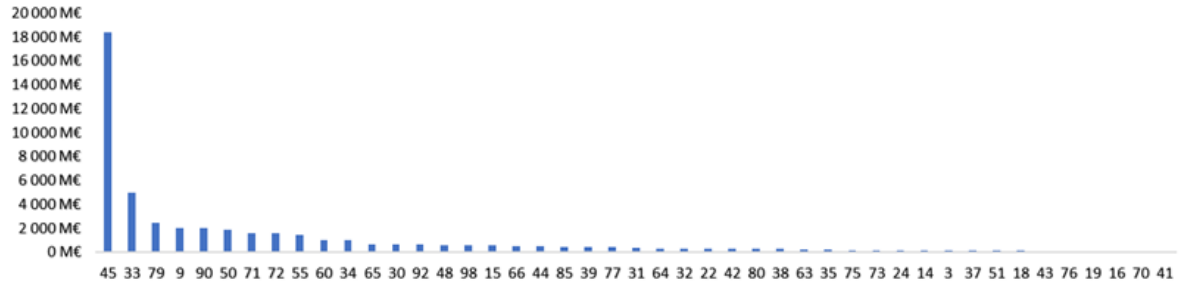


Figura 3 - Distribuição de valor por tipo de despesa (CPV's 2 Dígitos)

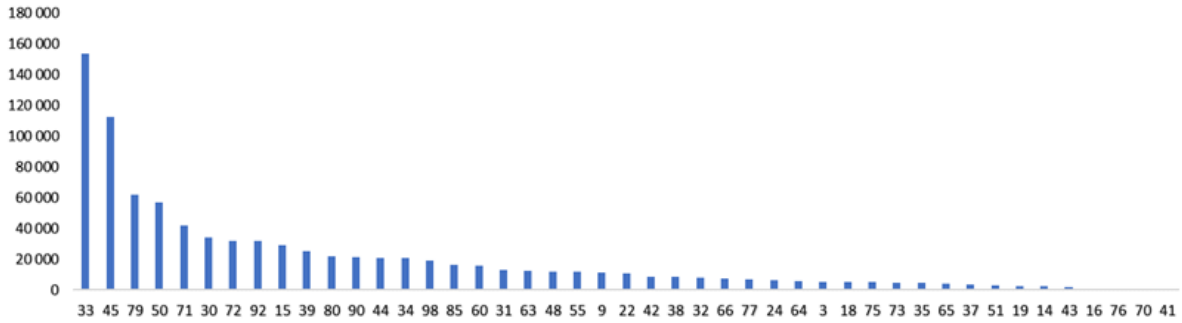


Figura 4 - Distribuição de nº de contratos por tipo de despesa (CPV's 2 Dígitos)

1.2.Objetivo do Projeto

O objetivo principal deste estudo é identificar clusters das **45** Divisões associadas ao código CPV (Dois primeiros dígitos), com desvios semelhantes face à relação entre despesa municipal e dimensão populacional ao nível **Município-Ano**. O conceito de desvio normalmente está associado ao conceito de anomalia, portanto podemos considerar que se pretende identificar as áreas de contratação onde os gastos anómalos se mostram mais ou menos intensos. A primeira fase do projeto consiste na extração dos contratos públicos do portal gov.pt com o uso de um web crawler desenvolvido em Python. A segunda fase consiste na transformação e limpeza da informação, neste caso, os contratos extraídos. Na terceira e última fase irá ser feita uma análise exploratória, a nível demográfico e a nível de distribuição, pelos diferentes tipos de contrato. Também é analisada a evolução temporal da relação entre o número de habitantes e montante associados aos contratos. Por fim é feita uma análise com recurso a técnicas de Data Mining para segmentar os tipos de despesa pública com base numa matriz de distâncias gerada pelos resíduos da relação entre **Despesa** e **Dimensão da População.**

1.3.Importância do Estudo

Os gastos de capitais públicos pelos governos locais tem sido motivo de diversos estudos. Alguns dos motivos pelos quais é importante medir a eficiência dos gastos locais deve-se ao facto de ser possível comparar essa mesma eficiência entre as diversas regiões (Farrel, 1957). Essa mesma comparação permite não só avaliar em detalhe o porquê da ineficiência da localidade em estudo bem como tomar medidas de ação por parte do governo local de forma a tornar as localidades mais eficientes ou até levar a novas eleições. (Lovell, 1993)

A eficiência de um Município pode ser medida de duas maneiras, uma incidente sobre o que consegue produzir em função dos recursos que tem (Output Direto) o que permite a posterior comparação de métricas entre os diversos locais. A outra que reside na satisfação e bem-estar da população (Output Consumidor). (Lovell, 2000)

Alguns setores importantes no estudo da eficiência de um Município são administração geral, educação, atividade social, saneamento e proteção ambiental. Estes setores podem ser traduzidos em variáveis concretas tais como população residente, residentes com mais de 65 anos, escolas per

capita, taxa de frequência escolar, uso de material reciclado, percentagem de população com água potável (Fisher, 1996).

Um estudo realizado nos Municípios de Lisboa e Vale do Tejo, usou para medição da eficiência dos contratos o método anteriormente descrito como Output Direto, ou seja, recursos/produção. Algumas das variáveis utilizadas são as mencionadas por Fisher. O estudo sugere que a mesma performance poderia ter sido alcançada usando 39% menos recursos, ou seja, a eficiência poderia ter sido superior sem gastos adicionais (Afonso & Fernandes, 2003).

O estudo elaborado nesta tese visa compreender que áreas de despesa dos Municípios Portugueses apresentam mais desvios através de clusters dos tipos de despesa pública criados a partir dos resíduos da relação entre a dimensão populacional e a despesa pública efetuada nos diversos setores definidos pelo primeiro nível de códigos dos CPV's (Divisões). Este estudo pode contribuir para detetar eventuais desvios anómalos em determinados tipos de despesa e inversamente aprendizagem sobre os grupos de despesas que menos desvios têm.

2.Metodologia e Ferramentas

2.1.Extração de dados

Atualmente existe uma proliferação de dados quer em dimensão quer em variedade. Os dados podem provir de diversos tipos de fontes, por exemplo, dados recolhidos utilizando aplicações informáticas, dados fornecidos por sensores ou dados de log associados a processamento de informação, por outro lado, podem estar organizados em bases de dados e apresentados em sites web.

A extração dos dados pode deparar-se com diversos problemas que podem ter a ver com a quantidade de dados, a qualidade dos dados e a disponibilidade dos dados. Os dados podem estar disponíveis de forma limitada quer em termos temporais quer em termos da capacidade de processamento necessária para a extração dos mesmos.

Relativamente à estruturação dos dados, estes podem estar estruturados numa base de dados relacional ou não estruturados como por exemplo em texto, imagens, áudio ou vídeo. Os dados poderão estar semi-estruturados como por exemplo em formato Extensible Markup Language (XML). (Saukar, Kedar, & Gode, 2018)

Quando se pretende criar uma "datawarehouse" a partir duma base de dados online tipo OLTP a extração de dados normalmente é o primeiro passo do chamado processo ETL (Extract, Transform and Load). Isto é, a extração é o passo antes da transformação e posterior carregamento numa "datawarehowse". A extração de dados é um processo que se repete ao longo do tempo para "alimentar" a "datawarehowse" em que são armazenados dados históricos agregados ou não de acordo com as necessidades de tratamento posterior e tendo em conta os limites de espaço assim como a rapidez de processamento. (Yamaganti & Sikharam, 2015)

Para a extração de dados a partir de bases de dados relacionais tipo OLTP a técnica de base é a utilização de qwery´s SQL. No entanto para outros tipos de fontes de dados tais comos dados em sites web é necessário desenhar programas específicos adaptados ao site Web em questão. Além disso, a automatização dessa extração também tem que ser específica pois depende da organização do site e dos métodos de pesquisa que o site põe à disponibilização do utilizador.

Como os dados que são tratados nesta tese estão disponíveis na web, utilizam-se técnicas de Web Scraping para a extração de dados a partir da Web sendo posteriormente tratados por algoritmos

estatísticos ou de Data Mining. As técnicas de Web Scraping são técnicas importantes para extrair dados a partir de sites web e transformá-los em dados estruturados. O objetivo básico do data-scraping é transformar a informação de web sites em estruturas compreensíveis tais como bases de dados ou ficheiros comma-separated values (CSV). Posteriormente esses ficheiros serão tratados e convertidos para os formatos adequados às ferramentas que irão ser utilizadas tais como ferramentas de análise estatística, de Machine Learning ou de Clustering. (Yamaganti & Sikharam, 2015)

A extração de dados a partir da web apresenta alguns desafios tais como automatizar o mais possível a extração de dados prescindindo o mais possível da intervenção humana, conseguir a adequada capacidade de processamento para extrair largos volumes de dados em relativamente pouco tempo, assegurar a privacidade dos dados em particular quando estão envolvidos dados pessoais. Quando se pretende utilizar algoritmos de Machine Learning para analisar os dados é necessário garantir a quantidade de dados suficiente para alimentar o processo de aprendizagem; maior exigência na manutenção dos programas de extração de dados pois poderão haver mutações da estrutura do web site ao longo do tempo. (Ferrara, Fiumara, & Baumgartner, 2012)

No caso dos sites web os dados são dum modo geral apresentados dentro de estruturas denominadas páginas web. As páginas web são documentos escritos em Hypertext Markup Language (HTML) e, mais recentemente, XHTML, que é baseado em Extensible Markup Language (XML). O objetivo do HTML é especificar o formato dos dados que serão interpretados pelos browsers web. Os documentos da web são representados por uma estrutura em árvore chamada Document Object Model. Existem várias técnicas de web scraping tais como: copiar e colar, programação HTTP (Hypertext Transfer Protocol), parsing de HTML (Hyper Text Markup Language), software Web Scraping, analisadores de páginas da web utilizando visão computacional. (Saukar, Kedar, & Gode, 2018)

A criação de programas para extração de dados a partir de sites é facilitada pela utilização de bibliotecas de software existentes para “crawling” e “scraping”. Um web crawler é um programa que navega de maneira metódica e automatizada na internet indexando as páginas da web. Scraping é uma técnica usada para ler informações de páginas da web com base em rotinas. (Vargiu & Urru, 2013)

2.2.Regressão linear

Quando temos na população de interesse duas variáveis x e y, chama-se modelo de regressão linear simples de duas variáveis ou bivariada ao modelo que relaciona as duas variáveis x e y através da equação

$$y = a + bx + u \quad (1)$$

que se supõe manter na população de interesse.

A variável y é a variável dependente e x é a variável independente. Uma análise de regressão trata efetivamente todos os fatores que possam influenciar y diferentes de *x* como não observados. Na equação (1) o parâmetro b é o declive na relação entre y e x, a o termo constante e u é denominado termo de erro ou perturbação no relacionamento, representando fatores diferentes de x que afetam y.

Dada uma amostra aleatória, o Método dos Mínimos Quadrados, é usado para estimar os parâmetros declive e constante da reta ajustada. Chamam-se valores ajustados aos valores $\hat{y}$ que, para cada abcissa observada x se encontram para sobre a reta ajustada. O resíduo duma observação é a diferença entre o seu valor real e seu valor ajustado i.e. $y - \hat{y}$. Uma boa estimativa para os erros aleatórios u é dada pelos resíduos.

O coeficiente de determinação, também chamado de R^2, é uma medida que varia entre 0 e 1, indicando a proporção da variação da amostra na variável dependente explicada pela variável independentes e serve como uma medida de qualidade do ajuste.

Nem sempre existem relações lineares entre variáveis dependentes e variáveis independentes, no entanto é possível incorporar muitas não linearidades na análise de regressão linear definindo adequadamente as variáveis dependentes e independentes. Algumas transformações comuns são *Lin-Log, Log-Lin e Log-Log* nas variáveis y e x respetivamente. (Wooldridge, 2013)

OUTLIERS

Outliers são os valores que estão muito longe, independentemente da direção, da média da distribuição de uma população. Outliers e pontos influentes podem ter um impacto significativo nos resultados de qualquer análise. Outliers não são necessariamente influentes nos coeficientes de regressão. Os pontos influentes devem ser tratados com o máximo cuidado, pois estes ao serem removidos, poderão levar a um modelo diferente. Outliers podem estar associados a outros pontos, por exemplo, na presença do primeiro Outlier, o segundo pode não funcionar como um Outlier. Ocorrência de Outliers pode ser por acaso. Se a ocorrência for por acaso, eles são descartados. (Dhakal, 2017)

Pode ser obtida uma medida para testar Outliers considerando a magnitude de cada resíduo, relativa ao seu desvio padrão a que se chama *Studentized Residual*. Pode obter-se outra medida para testar Outliers, removendo uma determinada observação da população obtendo-se o *Studentized Deleted Residual* (SDR). (Cousineau, 2010). Para um tamanho razoável da população, um SDR de magnitude 3 ou mais (valor absoluto) será considerado um Outlier. Qualquer um com magnitude 2-3 pode estar próximo de ser considerado Outlier dependendo do nível de significância usado.

Nem todos os Outliers têm uma forte influência sobre o modelo. Uma das medidas para detetar a influência de cada observação é a Distância de Cook, que considera a influência de um dado caso em todos os valores ajustados. Para avaliar se um dado é influente, Chatterjee indica uma Distância de Cook superior a 1 como uma diretriz operacional. (Chatterjee, 2000)

2.3.Clustering

Clustering refere-se à tarefa de agrupar dados semelhantes entre si de forma a criar grupos (clusters) distintos. Clustering é uma das técnicas usadas em Data Mining onde se procura descobrir padrões e relações entre dados. Além de clustering, Data Mining é uma área multidisciplinar que utiliza diversas áreas de conhecimento tais como tecnologias de bases de dados, integração de dados, transformação e limpeza de dados, machine learning e estatística. Como refere Xin-She Yang, "Data Mining expandiu o leque de técnicas para além das clássicas técnicas de modelação e métodos estatísticos".

Atualmente Data Mining é uma área com fortes contributos em diversas áreas como Saúde, Engenharia, Retalho, Banca, *Market basket Analysis* entre outras (Padhy N., 2012).

A compreensão de um conjunto de dados complexos é tipicamente facilitada pela agregação ou classificação dos objetos que compõem tais conjuntos em classes ou grupos. Estes identificam objetos que são entre si semelhantes, e que distinguem dos objetos remanescentes de forma idêntica. Assim, e de um modo geral os sistemas de classificação consistem na associação de um objeto a uma categoria, e à representação das relações ou estrutura entre as diferentes categorias. Os algoritmos de classificação podem dividir-se em algoritmos de tipo supervisionado ou não supervisionado.

No caso de algoritmos supervisionados, o ponto de partida é um conjunto de dados pré-classificados, devendo o algoritmo ser capaz de aprender com os dados existentes a classificar novos objetos dada as suas características de input. Por exemplo, a regressão logística ou as redes neuronais, são exemplos paradigmáticos de algoritmos de aprendizagem supervisionada.

No segundo caso dos algoritmos não supervisionados, não existem dados pré-classificados e por isso o objetivo é usar um algoritmo que dada a informação disponível retorne uma classificação/agregação possível dos dados existentes. Os dados, ou objetos, são assim aglomerados em classes, sendo os elementos pertencentes a cada classe formam um *cluster* cuja a identificação e caracterização é topicamente o objeto de estudo.

Os algoritmos de *Clustering* (DBSCAN, *K-Means*, *K-Medoids*, etc.) enquadram-se na classe de algoritmos de aprendizagem não supervisionada cujo objetivo é identificar a estrutura de um conjunto de dados não classificados através da organização dos dados em grupos homogéneos

onde a semelhança dentro do grupo é maximizada e a não semelhança fora do grupo também é maximizada. Cada um destes grupos assim formados designa-se por *Cluster*. *Clustering* utiliza diferentes métodos sendo os mais comuns os particionais e os hierárquicos. Os métodos particionais formam clusters mutuamente disjuntos e os hierárquicos criam uma estrutura em árvore. (Rui Xu, 2005).

A aplicação de algoritmos de *Clustering* é combinada usualmente com a definição de uma medida de distância e a definição duma função a otimizar. De seguida são apresentados os métodos que irão ser usados neste estudo.

ALGORITMO *K-MEANS*

O algoritmo *K-means* pretende criar uma partição dum conjunto de elementos criando K *clusters*. O algoritmo processa-se começando por selecionar K pontos iniciais a que se chamam centroides. De seguida formam-se K *clusters* associando cada elemento ao centroide mais próximo. Para cada *cluster* formado, calcula-se o valor médio dos seus pontos, passando esse valor a ser o novo centroide do *cluster*. Repete-se a formação de *clusters* utilizando como pontos iniciais os novos centroides. O processo termina quando os centroides não apresentam mudanças. Centroides são considerados os representantes dos Clusters.

Existem diversos algoritmos para implementar o *K-Means*, um dos mais conhecidos foi introduzido por McQueen (McQueen, 1967).

O algoritmo *K-Means* processa-se da seguinte maneira:

1. Iniciar aleatoriamente K centroides

2.Para cada ponto calcula-se a distância entre esse ponto e cada centroide, depois atribui-se o ponto ao centroide mais próximo, até todos os pontos estarem associados.

3. Recalcula-se o centroide de cada Cluster, calculando a média de todos os pontos do Cluster.

4. Repetir os passos 2-3 anteriores até que não haja alterações para cada cluster. (Rui Xu, 2005)

O objetivo do algoritmo *K-Means* é minimizar a soma quadrática dos erros:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||x_i^{(j)} - c_j||^2 \quad (2)$$

Onde $||x_i^{(j)} - c_j||^2$ é a distância entre o ponto $x_i^{(j)}$ e o centro de cluster c_j.

ALGORITMO HIERÁRQUICO

Os algoritmos Hierárquicos determinam *Clusters* que possuem *subClusters* – organizados em árvore. Cada *Cluster* é constituído pela reunião dos seus filhos. São produzidos agrupamentos hierárquicos começando com *Clusters* unitários e repetidamente aglutinando *Clusters* próximos dois a dois até chegar a um único *Cluster* com todos os elementos. Este percurso constitui uma estrutura hierárquica que graficamente se chama dendrograma.

A definição da distância entre *Clusters* define diferentes algoritmos. Um dos algoritmos é o "*single-linkage*" em que se considera como distância entre 2 *Clusters* a distância mínima que se pode encontrar entre dois elementos de cada um dos *Clusters*. O algoritmo "*complete linkage*" considera que a distância entre dois *Clusters* é o máximo das distâncias entre dois elementos de cada um dos *Clusters*. O algoritmo *"average-linkage"* utiliza como distância a média das distâncias entre cada par de membros de dois *Clusters*. Outro método designado por *"centroid-linkage"*, consiste em representar os *Clusters* pelo seu centroide tal como se faz no *K-Means* sendo a distância entre dois *Clusters* a distância entre os centroides. Outro método designado por "*group-average Clustering*" utiliza a média das distâncias entre todos os membros do *"merged Cluster"*. O método de *Ward* consiste em calcular o incremento entre a soma dos quadrados das distâncias para o centroide antes e depois de fundir os *Clusters* tentando minimizar este incremento em cada passo da iteração.

Todas esta medidas produzem a mesma estrutura hierárquica de *Clusters* se os *Clusters* forem compactos e bem separados, porém noutros casos podem produzir estruturas bastante diferentes. (Ian H. Witten)

ALGORITMO *K-MEDOIDS*.

O Algoritmo *K-Medoids* é análogo ao *K-Means* do ponto de vista que consiste em selecionar K pontos iniciais e depois formar K *clusters.*

A distância euclidiana é influenciada pelas maiores distâncias, levando a que o procedimento *K-Means* não tenha robustez contra os Outliers que produzem distâncias muito grandes. (Trevor Hastie, 2001). Nalguns casos é inviável remover todos os Outliers antes do processamento do algoritmo de *clusterização*, principalmente quando os dados são de alta dimensão. (Raykov, Boukouvalas, & Baig, 2016)

Foram propostas variações do *K-Means* que usam estimativas mais "robustas" para os centróides do *cluster*. Por exemplo, o algoritmo *K-Medoids* usa para centroide o ponto em cada *cluster* que está localizado mais centralmente. (Raykov, Boukouvalas, & Baig, 2016)

A única parte do algoritmo *K-Means* que assume a distância euclidiana é a etapa em que os representantes dos K *Clusters* são calculados como as médias dos respetivos *clusters*. Esta operação só pode correr sobre vetores em que se assumem variáveis (coordenadas) bem definidas. O algoritmo pode ser generalizado para uso com uma matriz de distâncias, substituindo o cálculo da média por uma otimização explícita para obter os centroides. Uma possibilidade é, em vez de fazer a média que gera um vetor que obviamente pode não pertencer à população, é restringir o centro de cada *cluster* a ser escolhido entre as observações atribuídas para o *cluster*. O Algoritmo *K-Medoids* segue esta estratégia, não dependendo de vetores perfeitamente definidos, pois desde que exista uma matriz de distâncias é possível encontrar o ponto mais central que pertencente à população do cluster.

O algoritmo de *Clustering K-Medoids* processa-se através dos passos seguintes (Trevor Hastie, 2001):

1. *Inicialmente selecionam-se K das n observações de dados como medoides e assigna-se cada observação ao medoide mais próximo, ficando assim formados K clusters.*
2. *Para cada cluster, encontra-se a observação no cluster que minimiza a distância total para os outros pontos desse cluster:*

$$m_k = \underset{\{x_i \in C_k\}}{argmin} \sum_{x_j \in C_k} D\ (x_i, x_j) \qquad (3)$$

logo o $\{m_k\}$*, para k = 1, 2,..., K são os novos centros dos clusters (medoides).*

3. *Dado o conjunto atual de medoides* $\{m_1, m_2, \dots, m_k\}$*, minimizar o erro total atribuindo cada observação ao medoide mais próximo, ficando assim formados novos K clusters.*
4. *Iterar os passos 2 e 3 até que os medoides não sejam alterados.*

Cada iteração para calcular um *cluster* provisório k no algoritmo *K-Means* requer uma quantidade de cálculo proporcional ao número de observações que lhe foram atribuídas, enquanto que no algoritmo *K-Medoids* a computação é muito mais intensiva.

MULTIDIMENSIONAL SCALING

A visualização de dados permite uma análise importante dos dados. Neste trabalho, como discutiremos mais à frente, trabalharemos sobre uma matriz de distâncias entre os objetos de análise. Recorreremos, portanto, ao *Multidimensional Scaling* (MDS) para inferir as posições de cada objeto num espaço n-dimensional que preserve a distância entre os objetos, mas que facilite a sua visualização num gráfico de dispersão. Outro método cujo objetivo é reduzir a dimensionalidade é o *Principal Component Analysis* (PCA) que determina as chamadas componentes principais através de vetores próprios. Neste estudo será usado o algoritmo MDS.

O algoritmo *Multidimensional Scaling* (*MDS*), pretende inferir a posição de cada objeto num espaço coordenado com menor número de dimensões. Assim, procura representar todos os objetos dados usando poucas dimensões e preservando a proximidade/distância entre os mesmos. (Rui Xu, 2005)

Considerando $\{x_1, x_2, \dots, x_n\} \in \mathbb{R}^p$, determina-se a distância $d(x_i, x_j\}$ entre pares de pontos x_i e x_j, pois o *MDS* utiliza apenas distâncias entre os pontos. Pretende-se encontrar uma representação dos pontos num espaço de menor dimensão preservando as distâncias entre os pares de pontos. O *MDS* pode inclusive prescindir dos pontos $\{x_1, x_2, \dots, x_n\} \in \mathbb{R}^p$ desde que já exista a matriz de

distâncias. O objetivo do *MDS* é encontrar $(z_1, z_2, \dots, z_n) \epsilon \mathbb{R}^k$ $(k < p)$ que minimizem a função de stress S. (Trevor Hastie, 2001):

$$S(z_1, z_2, \dots, z_n) = \sum_{i \neq j} (\mathrm{d}(x_i, x_j) - ||z_i - z_j||)^2 \quad (4)$$

TRATAMENTO DE DADOS E SEGMENTAÇÃO

O tratamento de dados é comum e necessário para aumentar a capacidade dos modelos preditivos e de *Clustering* de extrair informações úteis. Para colmatar erros e omissões nos dados podem utilizar-se diversas abordagens tais como imputações de valores ausentes, suavização para remover o ruído sobreposto ou excluir os exemplos discrepantes. Podem utilizar-se transformações de variáveis, de escala e centralização dos valores dos dados e até métodos mais avançados. (Xi Hang Cao, 2016)

Na fase de pré-processamento e pós-processamento, o recurso seleção / extração (bem como padronização normalização) e validação de *Cluster* são tão importantes como os algoritmos de agrupamento. Infelizmente, ambos os processos não têm orientação universal. (Rui Xu, 2005)

Um método simples e amplamente utilizado para estimar a distância entre observações com alguns componentes ausentes é a Partial Distance Strategy (PDS). Na PDS, uma estimativa para o quadrado da distância entre observações é encontrada calculando a soma das diferenças quadráticas dos componentes mutuamente conhecidos e dimensionar o valor proporcionalmente para os valores ausentes. (Eirola E, 2013)

A normalização antes do *Clustering* é especificamente necessária quando se utiliza distâncias como a distância euclidiana que é sensível a variações na magnitude ou nas escalas dos atributos. (Usman, 2013). Técnicas de normalização de dados incluem normalização *min-max*, normalização *Z-Score* e normalização da escala decimal. Não há regras definidas universalmente para normalizar conjuntos de dados e, portanto, a escolha de uma regra de normalização específica é largamente deixada ao critério do utilizador (Usman, 2013). O método de normalização *RobustScaler* usa um método semelhante ao *min-max*, mas usa o intervalo interquartil, para que seja robusto para valores extremos. (Keen, 2019)

AVALIAÇÃO DE *CLUSTERS*

Pretende-se uma medida que se possa aplicar a um conjunto de *clusters* (partição) de modo a poder comparar dois conjuntos de *clusters* (partições) correspondentes a diferentes análises. O Coeficiente *Silhouette* é uma medida da qualidade do *clustering* que combina ideias de coesão e separação.

Uma das vantagens do Coeficiente *Silhouette* é ser capaz não só de avaliar a qualidade do Clustering formado por um determinado algoritmo como ser útil no que toca à escolha do número de clusters. Provavelmente o método mais conhecido seja o "método do cotovelo" conhecido como "elbow-method". O "método do cotovelo" tenta minimizar a soma das distâncias intra-cluster,, enquanto o método Silhouette avalia não só a distância intra-cluster como a distância inter-cluster tornando esta medida mais robusta. Kaufman and Rousseeuw afirmam que o método do cotovelo é às vezes ambíguo e que o método *Silhouette* pode ser uma alternativa para qualquer método de avaliação de *Clustering*. Outra vantagem do método *Silhouette* é ser mais claro visualmente na escolha do número de clusters já que para cada K cluster's apresenta um máximo para o melhor K. O Coeficiente *Silhouette* geralmente varia entre 0 e 1, sendo considerado melhor se mais próximo de 1. (Santhana Chaimontree, 2010). Os dois gráficos seguintes mostram a diferença entre os dois métodos para um conjunto de pontos aleatórios num espaço de 4 dimensões[3]. A **Figura 6** relativa ao método *Silhouette* permite observar que o número de clusters indicado é 5 onde existe um máximo visível. Já o método do cotovelo (**Figura 5**) não é tão claro a indicar se a escolha do número de clusters deverá ser 4, 5 ou 6. Este método consiste em escolher o número de clusters onde o gráfico apresenta uma maior quebra repentina (cotovelo) o que neste caso não é evidente.

[3] https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6

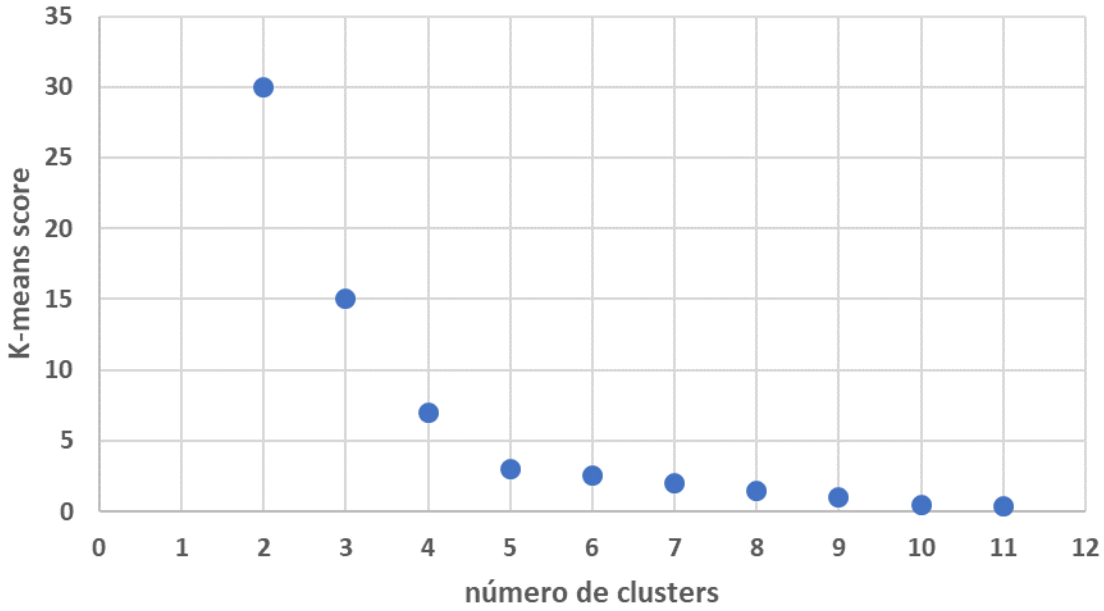


Figura 5 - Número de clusters (Método Cotovelo)

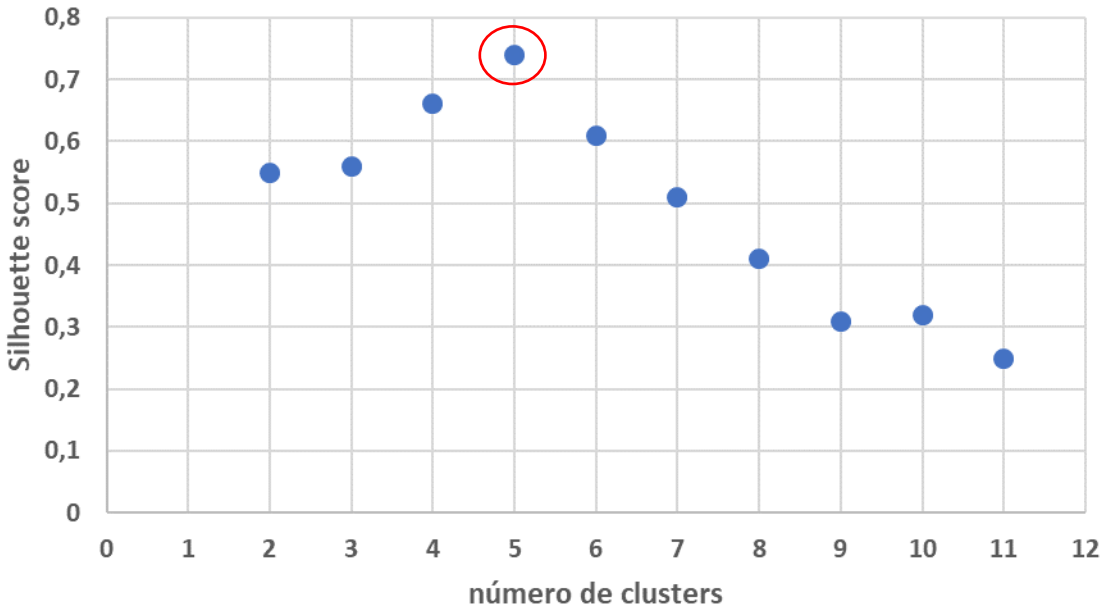


Figura 6 - Número de clusters (Silhouette)

Neste projeto irá ser usado o método Silhouette para avaliação dos clusters. O cálculo do Coeficiente *Silhouette* processa-se em duas fases, numa primeira fase determina-se o Coeficiente *Silhouette* para cada ponto individualmente e numa segunda fase o coeficiente geral para o conjunto de *clusters.* (Santhana Chaimontree, 2010)

2.4.Python e Bibliotecas

Neste trabalho a recolha dos dados e respetivo processo foram executados em Python. A escolha desta linguagem de programação deveu-se à sua popularidade e recursos disponíveis, nomeadamente no que respeita às bibliotecas para executar as tarefas de *WebScraping*, *Data Cleanning e Clustering.*

Python é uma linguagem de programação de alto nível, que permite trabalhar áreas como desenvolvimento de aplicações *web* e *DataScience*. É adequada para quem trabalha em áreas relacionadas com *Big Data* e *Analytics* devido à elevada capacidade de processamento e ao facto de possuir inúmeras bibliotecas relacionadas com Análise Estatística e *Machine Learning*. Neste projeto, foram utilizadas várias destas bibliotecas. Em particular foi utilizada a biblioteca ***Beautiful Soup*** cuja finalidade é fornecer ferramentas para a extração de dados *web*. Foram também utilizadas as bibliotecas ***Statsmodels*** e ***Scikit-learn*** para análise estatística e de *Clustering***.** Na tabela seguinte são apresentadas as bibliotecas/módulos que foram usadas durante este projeto.

Biblioteca	Descrição
pandas	Biblioteca usada para manipulação de tabelas
NumPy	Biblioteca usada para manipulação de *Arrays*
SciPy	Biblioteca de funções matemáticas
matplotlib	Biblioteca para impressão de gráficos
sklearn	Biblioteca para "machine learning"
itertools	Módulo que fornece funções para construção de iterações
random	Módulo que implementa geradores de números aleatórios
math	Módulo que fornece funções matemáticas
Beautiful Soup	Biblioteca que fornece um parser de HTML/XML
pyclustering	Biblioteca de data mining

Tabela 1 - Bibliotecas/Módulos utilizados

3.Execução e Análise

O tratamento e a análise dos dados deste estudo podem ser separados em três fases: 1) **Extração**; 2) **Transformação**; e 3) **Análise.**

Extração correspondeu ao processo de extração automática dos dados existentes no portal base, http://www.base.gov.pt/. Os dados correspondem à informação sobre os contratos realizados entre instituições públicas e privadas entre 2008 e 2017 em Portugal. Para a recolha da informação foi desenvolvido um *crawler* usando a linguagem de programação *Python*. Adicionalmente, foram extraídos dados sociodemográficos referentes aos municípios, estes dados foram extraídos do portal PorData (https://www.pordata.pt). Essencial para o entendimento da informação em cada contrato, foi necessário obter informação detalhada sobre a classificação CPV (Common Procuremment Vocabulary). Esta informação foi extraída de https://www.espap.pt/. **Transformação** correspondeu ao tratamento dos dados em bruto para a obtenção de dados que fossem de encontro aos objetivos pretendidos para este estudo. Tudo o que é referente a de limpeza dos dados e junção de mais informação aos contratos como quais dos contratos são referentes a empresas municipais e a dimensão populacional dos municípios associados.
Análise é a última fase que consiste em segmentar as diferentes tipologias de contrato, neste caso de classes de CPV's através de diferentes métodos de *Clustering* como o ***Hierárquico***, ***K-Medoids*** e ***K-Means*** usando a matriz de distâncias construída através dos resíduos da relação entre **Despesa** e **Dimensão População.**

A imagem seguinte (**Figura 7**) ilustra o processo e as 3 diferentes fases mencionadas anteriormente.

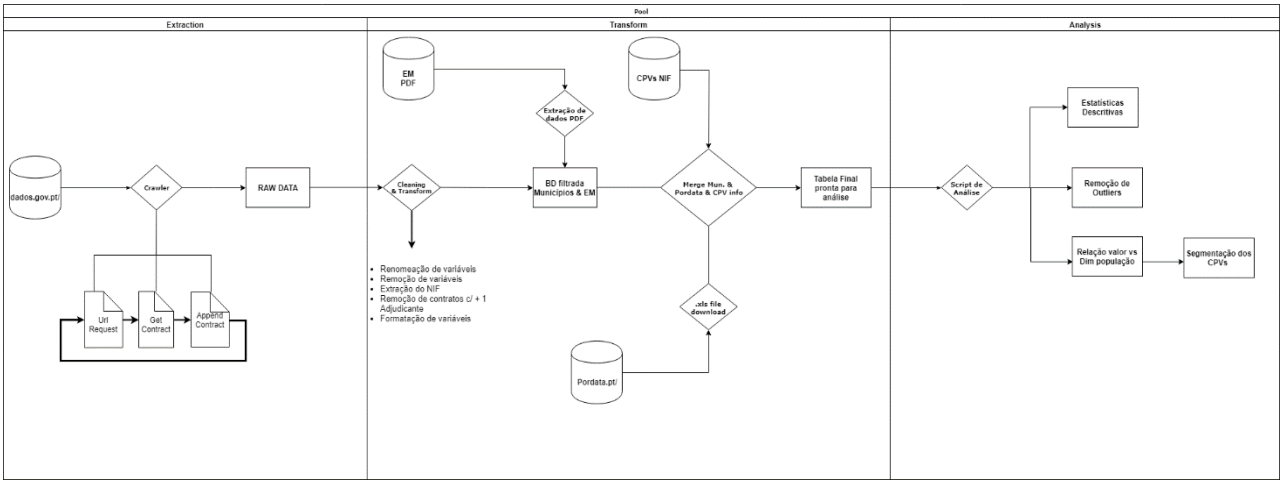


Figura 7 - Diagrama Global de Extração, Transformação e Análise (ETA) dos dados.

3.1.Extração de dados web

A extração dos dados do portal Base.gov.pt foi feita via a implementação de um *web crawler* desenvolvido em ***Python*** e com recurso à biblioteca *Beautiful Soup*. Os contratos foram extraídos em *batchs* anuais para evitar problemas de sobrecarga de pedidos e bloqueio da página. O modo como o *crawler* foi desenvolvido teve por lógica a de carregar o menos possível o servidor com pedidos. Desse modo, foi dividido em dois passos. No primeiro foram descarregados todos os links de cada contrato, e em segundo a lista final de links obtida foi iterada para extrairmos a informação de cada contrato específico. Posto isto foi necessário desenvolver duas funções principais, uma de *request* do contrato link a uma segunda função dedicada à extração do contrato. O diagrama da **Figura 8** demonstra o *workflow* descrito. As três fases que consistem no são detalhadas nas secções seguintes.

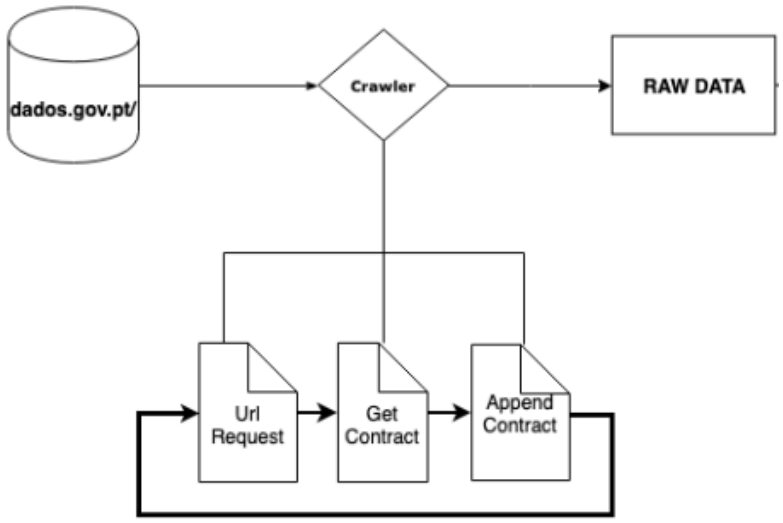
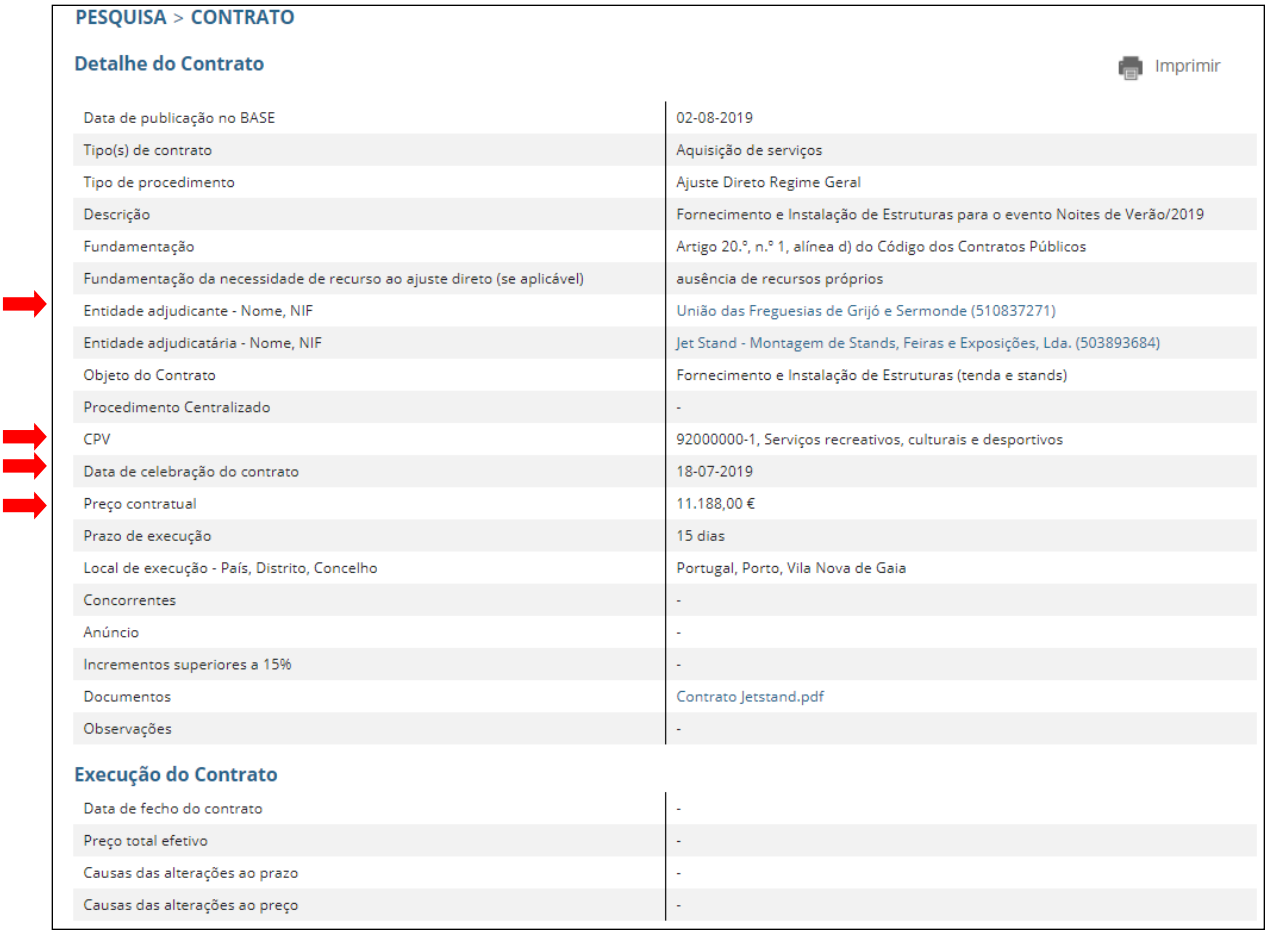


Figura 8 - Diagrama da Extração

A **Figura 9** exemplifica a informação de um contrato como está disponibilizada no portal base.gov.pt. Alguns contratos têm campos um campo extra aos que estão ilustrados na figura, mas que não eram importantes para este estudo. Os campos que iram ser usados durante a análise são identificados com uma seta vermelha.



PESQUISA > CONTRATO

Detalhe do Contrato

Imprimir

Data de publicação no BASE	02-08-2019
Tipo(s) de contrato	Aquisição de serviços
Tipo de procedimento	Ajuste Direto Regime Geral
Descrição	Fornecimento e Instalação de Estruturas para o evento Noites de Verão/2019
Fundamentação	Artigo 20.º, n.º 1, alínea d) do Código dos Contratos Públicos
Fundamentação da necessidade de recurso ao ajuste direto (se aplicável)	ausência de recursos próprios
Entidade adjudicante - Nome, NIF	União das Freguesias de Grijó e Sermonde (510837271)
Entidade adjudicatária - Nome, NIF	Jet Stand - Montagem de Stands, Feiras e Exposições, Lda. (503893684)
Objeto do Contrato	Fornecimento e Instalação de Estruturas (tenda e stands)
Procedimento Centralizado	-
CPV	92000000-1, Serviços recreativos, culturais e desportivos
Data de celebração do contrato	18-07-2019
Preço contratual	11.188,00 €
Prazo de execução	15 dias
Local de execução - País, Distrito, Concelho	Portugal, Porto, Vila Nova de Gaia
Concorrentes	-
Anúncio	-
Incrementos superiores a 15%	-
Documentos	Contrato Jetstand.pdf
Observações	-

Execução do Contrato

Data de fecho do contrato	-
Preço total efetivo	-
Causas das alterações ao prazo	-
Causas das alterações ao preço	-

Figura 9 - Exemplo de contrato[4] e identificação de campos utilizados para análise

[4] http://www.base.gov.pt/Base/pt/Pesquisa/Contrato?a=5750760

De seguida é feito um breve resumo das três principais funções que compõe o web crawler: ***Url Request***, ***Get Contract*** e ***Append Contract***

URL REQUEST

```
import datetime
#import time

def Url_NextTimeGap (dataInicio, dataFim, Dias): #Parametros de entrada (DataIncio e DataFim)
    dataInicio = datetime.datetime.strptime(dataInicio, "%Y-%m-%d") #Transformar DataInicio para formato Data
    dataInicio_Date = dataInicio + datetime.timedelta(days=Dias) #Adiciono 15 dias à Data Inicio

    dataFim = datetime.datetime.strptime(dataFim, "%Y-%m-%d") #Transformar DataFim para formato Data
    dataFim_Date = dataFim + datetime.timedelta(days=Dias) #Adiciono 15 dias à Data Fim

    #Passo as data de novo para Str para poder mudar o url para um novo intervalo de tempo

    dataInicio=dataInicio_Date.strftime('%Y-%m-%d') #DataInicio para Str
    dataFim = dataFim_Date.strftime('%Y-%m-%d') #DataFim para Str

    #Definição do novo url com novo intervalo de tempo

    url="http://www.base.gov.pt/Base/pt/ResultadosPesquisa?type=contratos&query=texto%3D%26tipo%3D0%26tipocontrato%3D0%26cpv%3D%26n
        + "desdedatacontrato%3D"+dataInicio+"%26atedatacontrato%3D"+dataFim \
        + "%26desdedatapublicacao%3D%26atedatapublicacao%3D%26desdeprazoexecucao%3D%26ateprazoexecucao%3D%26desdedatafecho%3D%26ate

    return url, dataInicio_Date, dataFim_Date, dataInicio, dataFim #Retornar novoUrl, datas em formato Data e formato Str
```

Figura 10 - Função *URL Request*

Função que atualiza o *request* do *url* onde se encontram os contratos relativos a 1 dia. Esta função recebe como argumentos iniciais:

DataInicio (*text*) **–** Data em que começa a extração de contratos. Esta data é inicialmente transformada em formato *Data* para que possa ser incrementada do parâmetro *Dias* e de seguida volta a passar para formato texto *YYYY-MM-DD* para ser concatenada na *string* que forma o próximo link de *request*

DataFim (*text*) **–** Data em que acaba a extração de contratos. Esta data é inicialmente transformada em formato *Data* para que possa ser incrementada do parâmetro *Dias* e de seguida volta a passar para formato texto *YYYY-MM-DD* para ser concatenada na *string* que forma o próximo link de *request*

Dias (*int*) **–** Número de dias que pretendo carregar entre a DataInicio e a DataFim. Este parâmetro foi criado por uma questão de performance e peso para o servidor. Tem o objetivo de escolher quantos dias entre a DataInicio e a DataFim queremos carregar de uma vez no servidor isto é: se escolher para a Datainicio e DataFim 1/1/2009 e 15/1/2009 respetivamente, esta opção se for **14** a página vai carregar todos os contratos entre estas datas apenas com um link o que vai ser extremamente massivo para o servidor. Eu optei por ter esta opção sempre a **1** carregando apenas os contratos dia a dia, ou seja, o ***Url_Request*** vai devolver 14 links diferentes.

A atualização do link baseia-se na data parametrizada no início do *scrip*. Este link é formado através da concatenação entre as DataInicio, DataFim e a restante *string* que compõe o link.

GET CONTRACT

Função responsável pela extração do respetivo contracto público. Esta função utiliza as bibliotecas do *Python* **Beatuifull Soup e Requests** próprias para trabalhar com extração de dados da *web*. Esta função recebe como argumento o *url* que encaminha para a página do contrato. Durante o processo de recolha dos contratos existiram diversos problemas ao nível de conexão com o servidor e links quebrados o que impedia fazia constantemente o processo parar. A solução foi criar um sistema de *debbugging* de forma a perceber identificar o eventual problema quando se acede ao link do contrato. São feitas no máximo **15** tentativas de ligação num timing máximo de **60** segundos como mostra a imagem seguinte.

```
from bs4 import BeautifulSoup
import requests
import time
import pandas as pd
pd.options.mode.chained_assignment = None

def ContratoData (url):
    print(url[51:])
    tries=0
    exception = True
    while (exception and tries < 15) :
        try:
            html = requests.get(url, timeout = 60)
            if html.status_code==500:
                tries=tries+1
                html.raise_for_status()
            exception = False
                #print("ok")
        except requests.exceptions.HTTPError:
            print("Error 500, Reconnecting..")
            print(tries)
            time.sleep(1)
            exception = True
            #print("ok")
        except requests.ConnectionError:
            print("Conection Error, Reconnecting..")
            time.sleep(30)
            exception = True
        except requests.RequestException:
            print("Handling RequestException, Reconnecting..")
            time.sleep(30)
            exception = True
        except requests.Timeout:
            print("TimeOut Error , Reconnecting..")
            time.sleep(30)
            exception = True
```

Figura 11 - Função *GetContract*

requests.exceptions.HTTPError: - Deteta erros do tipo 500 isto é, se o link associado ao contrato está quebrado ou simplesmente não existe. Este erro é o mais grave, porque não tem resolução independentemente do número de tentativas.

requests.ConnectionError: - Deteta erros de conexão quer seja do lado servidor quer seja do lado de onde esteja a ser executado o programa

requests.RequestException: - Deteta se o *request* feito não respondeu temporariamente. Normalmente este tipo de erro funciona após algumas tentativas de novo *request*.

Após a fase de *debug* segue-se a o processo de *storing* do respetivo contrato. A página *web* onde este está guardado tem tipicamente 2 tabelas onde guarda a informação. Para as detetar, é utilizado o método bs.*find* que procura todas as tabelas na página. Como estas tabelas são as únicas do tipo 'tabela' na página *web*, basta guardar o seu conteúdo num D*ataFrame* de uma única coluna através do método *append* e de seguida fazer a transposição do mesmo para ficar com um DataFrame de uma linha e 28 colunas. Além das colunas que compõem o contrato ainda são adicionadas duas colunas extra, '**Status**' com o valor '**OK**' ou '**ERRO**' para os contratos que passaram ou não passaram na fase de debug respetivamente e '**ID**' com o número do contrato (últimos 8 dígitos da *string* de cada link).

APPEND CONTRACT

Script principal do programa onde se desenrola todo o processo entre as datas parametrizadas entre as quais queremos fazer a extração e a respetiva agregação dos contractos. É composto por dois ciclos *'while'*, o primeiro responsável por percorrer todos os dias entre as datas parametrizadas 1 a 1. O segundo ciclo *'while'* é responsável por percorrer todas as páginas de contratos que compõem um dia. Cada página com os links dos contratos permite um acesso mais detalhado ao mesmo através de outro link representado pelo símbolo *'+'*. Estes links *('+')* estão todos dentro de uma tabela que é única na página e é detetada através do método *bs.find('table')* e posteriormente detetados através do teste *tr.find('a').getText()== '+'*. Após aceder a este link é invocada a função *GetContract* onde é feito o append do contrato. Após fazer o append de todos os contratos em cada página é procurado o link da *'PáginaSeguinte'* e repetir o processo. Este 'botão' é detetado através do método *bs.find_all('a', {'class': 'prev'})*. Quando é encontrada uma página sem contratos (critério de paragem do primeiro *while*) é altura de definir novo intervalo de

tempo isto é andar um dia para a frente e recomeçar o primeiro *while*. Quando a nova data é igual à DataFim o processo acaba.

3.2.Transformação de dados

Após a extração dos contratos é feito todo o trabalho de limpeza e manipulação da *RawData*. Ainda é feita a junção de novos dados de diferentes fontes de modo a identificar que contratos são referentes a empresas municipais. Por fim ainda é adicionada informação referente à dimensão do município onde o contrato foi celebrado. O diagrama seguinte mostra fluxo deste processo:

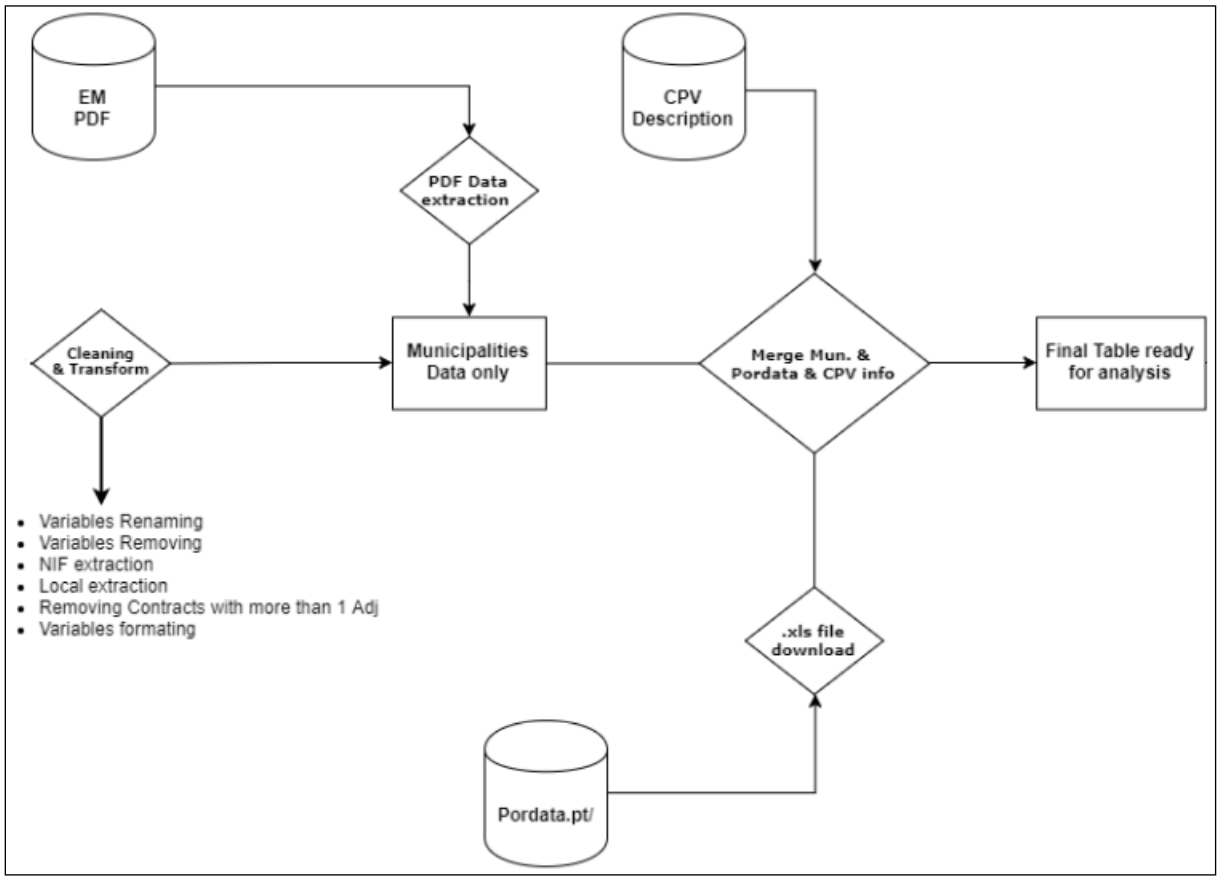


Figura 12 - Diagrama Transformação de dados

A tabela seguinte mostra a estrutura da base de dados após extração e respetiva mudança de nome das variáveis. As principais transformações efetuadas nestes campos são descritas a seguir à **Tabela 2** - Campos após extração do site base.gov.pt

.

Campo	Campo Renomeado	Tipo
Data de publicação no BASE		*Str*
Tipo(s) de contrato		*Str*
Tipo de procedimento		*Str*
Descrição		*Str*
Fundamentação		*Str*
Fundamentação da necessidade de recurso ao ajuste direto (se aplicável)		*Str*
Entidade adjudicante - Nome, NIF	**contracting**	*Str*
Entidade adjudicatária - Nome, NIF	**contracted**	*Str*
Objeto do Contrato		*Str*
Procedimento Centralizado		*Str*
CPV	**CPV**	*Str*
Data de celebração do contrato	**date**	*Str*
Preço contratual		*Str*
Prazo de execução		*Str*
Local de execução - País, Distrito, Concelho		*Str*
Concorrentes		*Str*
Anúncio		*Str*
Incrementos superiores a 15%		*Str*
Documentos		*Str*
Observações		*Str*
Data de fecho do contrato		*Str*
Preço total efetivo		*Str*
Causas das alterações ao prazo		*Str*
Causas das alterações ao preço		*Str*
ID		*int*
Status	**status**	*Str*
Nº de registo do acordo quadro		*Str*
Descrição do acordo quadro		*Str*
Convidados		*Str*

Tabela 2 - Campos após extração do site base.gov.pt

A transformação principal no que toca a variáveis é efetuada no campo '*contracting'* onde o objetivo é isolar o **NIF** do respetivo adjudicante associado. Devido à possibilidade de haver mais que uma entidade adjudicante a contratar foi necessário criar um dicionário com a ajuda da seguinte função:

```
#split Names and returns dictionary
def splitNames(x):
    if x == "-": #if empty return empty
        return x;

    temp = x  #save string in temp
    tnif = extractNIF(x) # extract NIF to list
    output = []  #create output variable, surprise it will be list of dictionaries

    if len(tnif) > 1: #using tnif check if there are more than one entity in the list
        #lets us use the nif to split the strings, there might be better ways of doing this
        for s in tnif:
            temp = temp.replace('('+str(s)+')','('+str(s)+');')
        temp = temp[:-1]
        temp = temp.split(';')

        for i in range(len(temp)):
            output.append({'Name':name(temp[i], get(tnif,i)), 'NIF': get(tnif,i)})
    else: #if there is only one element then operation is trivial
        output.append({'Name': name(temp, get(tnif,0)), 'NIF':get(tnif,0)})
    return output
```

Figura 13 - Função Isolamento do NIF

Exemplo do campo '*contracting*' com duas entidades adjudicantes após transformação:

[{'Name': 'Epr Viana do Castelo', 'NIF': '600003728'}, {'Name': 'Direcção-Geral dos Serviços Prisionais', 'NIF': '600000117'}]

A segunda transformação necessária para fazer análise exploratória inicial foi mudar o campo *'value'* que ao ser de tipo *'Str'* não permitiria qualquer tipo de agregação da despesa. Esta transformação foi feita através da função ilustrada na imagem seguinte:

```
df3['value'] = df3['value'].map(lambda x: float(x[0:-5].replace(".","")+'.'+x[-4:-2])) # convert value to number
```

Figura 14 - conversão variável *'value' para 'float'*

A última transformação para a fase pré-exploração é feita na variável *'date'* onde foram construídas as variáveis *'Ano', 'Mês', 'Dia' através da função*

```
df3['Year']= df3['date'].apply(lambda x: x.split("-")[2]) #Separate Year
df3['Month']= df3['date'].apply(lambda x: x.split("-")[1]) #Separate Month
df3['Day']= df3['date'].apply(lambda x: x.split("-")[0]) #Separate Day
```

Figura 15 - Função para separação de data em Ano, Mês e Dia

3.3.Limpeza de registos

Após a primeira fase de transformação das variáveis acima mencionadas é possível começar a fazer alguma exploração. De seguida foi utilizado a variável *'Status'* referente a contratos com link quebrado durante a **Extração** para perceber a dimensão dos mesmos. Como podemos ver na **Tabela 3** em baixo apenas existem 9 contratos com *'status'* errado, ou seja, uma percentagem ínfima da base pelo que serão eliminados ficando assim com **883.726** contratos.

status	Contratos	%
OK	883.726	99,999%
ERRO	9	0,001%

Tabela 3 - % Contratos c/ Links quebrados

Alguns dos contratos presentes na base não têm CPV associado impossibilitando os mesmos de serem considerados para a análise. Como se pode ver na **Tabela 4** estes contratos também representam uma percentagem quase nula (apenas **71**) sendo assim eliminados ficando assim **883.655** contratos presentes.

CPV status	Contratos	%
CPV OK	883.655	99,992%
CPV	71	0,008%

Tabela 4 - % Contratos s/ CPV

Ainda relativamente a contratos com *missing values* há registos sem adjudicante associado. Estes contratos também terão de ser eliminados. A tabela seguinte mostra que apenas se tratam de **7** contratos. Como se pode visualizar também se trata de uma percentagem pequena pelo que são eliminados estes registos (**Tabela 5**).

Contracting	Contratos	Contratos
contracting OK	883.648	99,999%
contracting	7	0,001%

Tabela 5 - % Contratos s/adjudicante

Como referido. anteriormente um contrato pode ter dois ou mais adjudicantes associados. Para poder perceber a dimensão dos contratos associados a mais de um adjudicante foi necessário criar uma coluna adicional '*CountNIF'* e agrupar por este campo.

A figura seguinte (**Figura 16**) mostra o quão residual são os contratos associados a mais de um adjudicante.

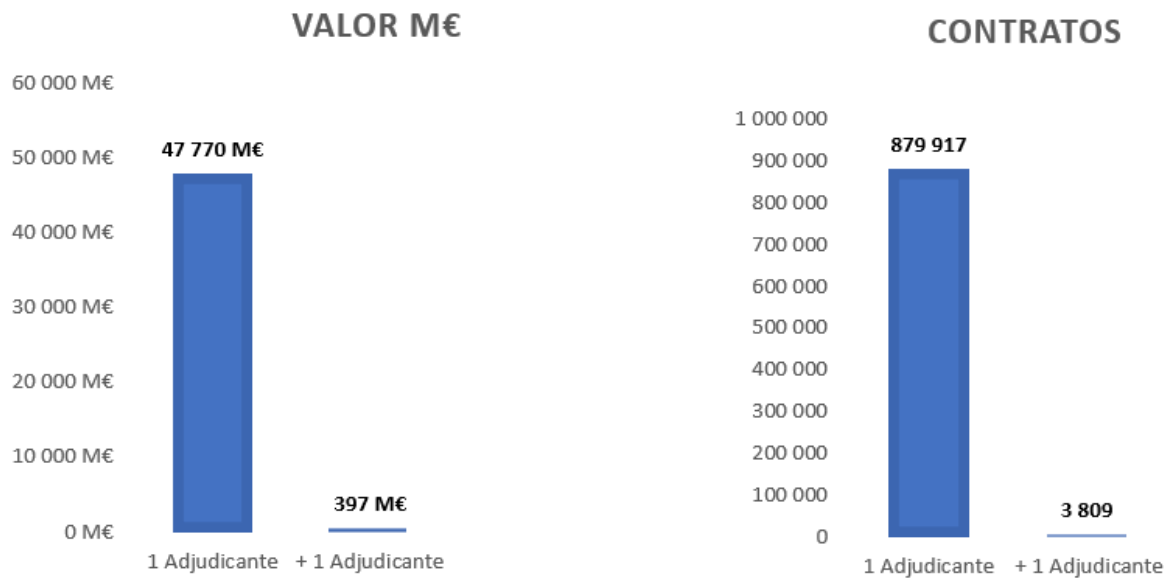


Figura 16 – Despesa e nº de contratos associados a mais de 1 adjudicante

Em função dos resultados obtidos em cima consideramos que seria justo trabalhar apenas com os contratos associados apenas a um adjudicante. Esta consideração facilitou bastante a associação entre um contrato e um adjudicante. Caso tivesse considerado trabalhar com toda a base teria um problema de como seria dividida a despesa entre os diferentes adjudicantes sendo que apenas há um valor registado no contrato público. Após a eliminação ficaram **879.839** registos.

IDENTIFICAÇÃO DOS CONTRATOS ASSOCIADOS A MUNICÍPIOS/EMPRESAS MUNICIPAIS

O objetivo desta fase é encontrar identificar todos os contratos associados a municípios e empresas municipais dos **879.839**. Como já referenciado, cada contrato tem um **NIF** associado ao adjudicante. Sendo o objetivo deste trabalho estudar a despesa pública associada aos municípios foi necessário identificar não só os municípios como também as respetivas empresas municipais o que em conjunto totalizará a despesa total associada a um determinado município. A identificação do **NIF** dos municípios[5] foi extraída da internet e posteriormente anexado a um ficheiro *.xls.* A identificação das empresas municipais foi extraída ficheiros públicos em formato *pdf do site https://www.occ.pt/* [6]. Estes ficheiros apenas contêm informação das EM que realizaram despesa, ou seja, pode haver EM em apenas alguns ficheiros. No Anexo **2**é possível ver um exemplo do ficheiro *pdf* de 2017.

Como podemos ver no **Anexo 2**, Almada é o primeiro município e tem duas despesas associadas: Serviços Municipalizados De Água E Saneamento (**SMAS**) e ECALMA-Estacionamento e Circulação (**EM**).

De forma a ficar apenas com contratos associados a Municípios e a EM foram feitas duas operações. A primeira entre a tabela de Municípios/NIF (**Anexo 3**) e a tabela de contratos com *Left Join* no campo **NIF**. A segunda entre a tabela Municípios/EM (**Anexo 4**) e a tabela de contratos de novo com *Left Join* no campo **NIF**. De seguida é feito o *append* entre estas 2 tabelas.

A figura seguinte (**Figura 17**) ilustra o processo descrito anteriormente.

[5] https://codigopostal.ciberforma.pt/

[6] https://www.occ.pt/pt/a-ordem/publicacoes/anuario-financeiro-dos-municipios-portugueses/.

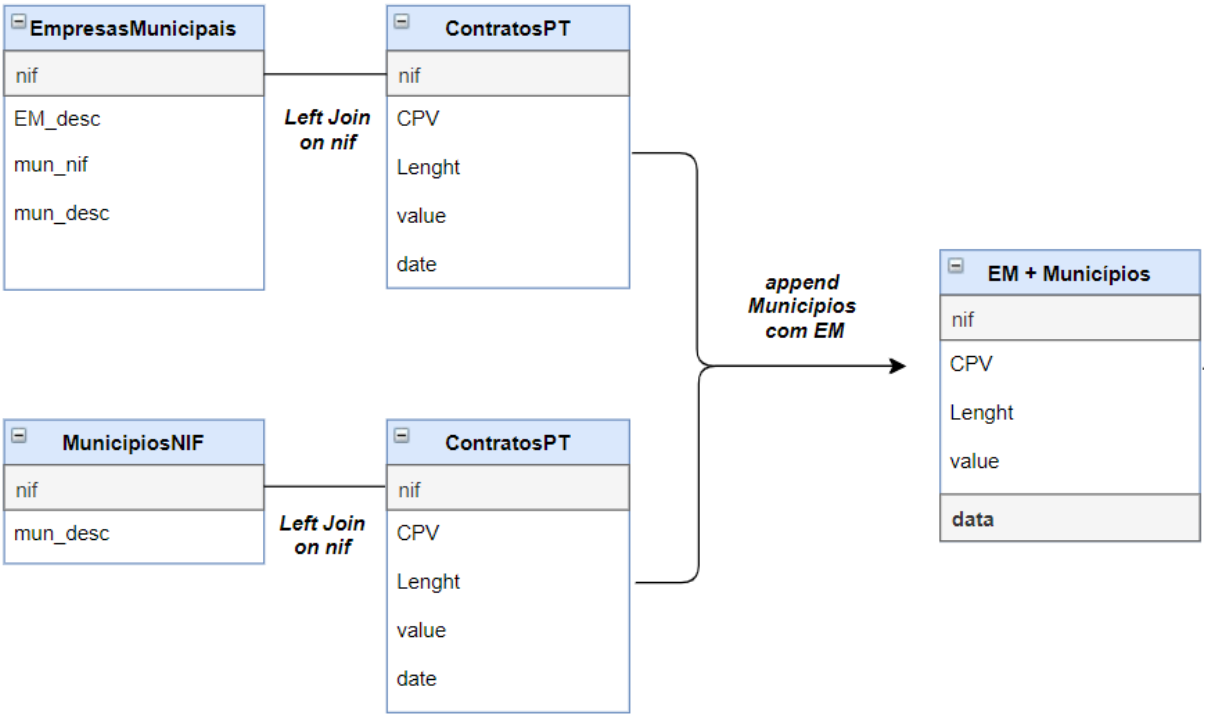


Figura 17 – Diagrama relativo à criação de tabela de contratos referente a municípios e EM

Como a **Figura 17** mostra apenas ficamos com algumas variáveis das que tínhamos inicialmente extraído.

Como um dos objetivos deste projeto é estudar a evolução anual dos municípios a nível de despesa ou da relação despesa/população escolhemos trabalhar só com os dados entre **2009** e **2017** devido ao facto de o ano de 2018 ainda não estar completo no momento em que feita a extração dos dados. O número de registos final foi de **303.449** sendo **225.794** (**74%**) relativos a municípios e **77.655** (**26%**) a empresas municipais. Neste ponto ficamos com apenas **34%** dos **883.735** contratos inicialmente extraídos.

MERGE DE INFORMAÇÃO RELATIVA A MUNICÍPIOS E CPV'S

É nesta etapa que é feita a junção de informação de outras fontes que não estão presentes na base de dados principal nem no site de onde foram extraídos os contratos, **base.gov.pt.**

A figura seguinte mostra o fluxo deste processo (**Figura 18**):

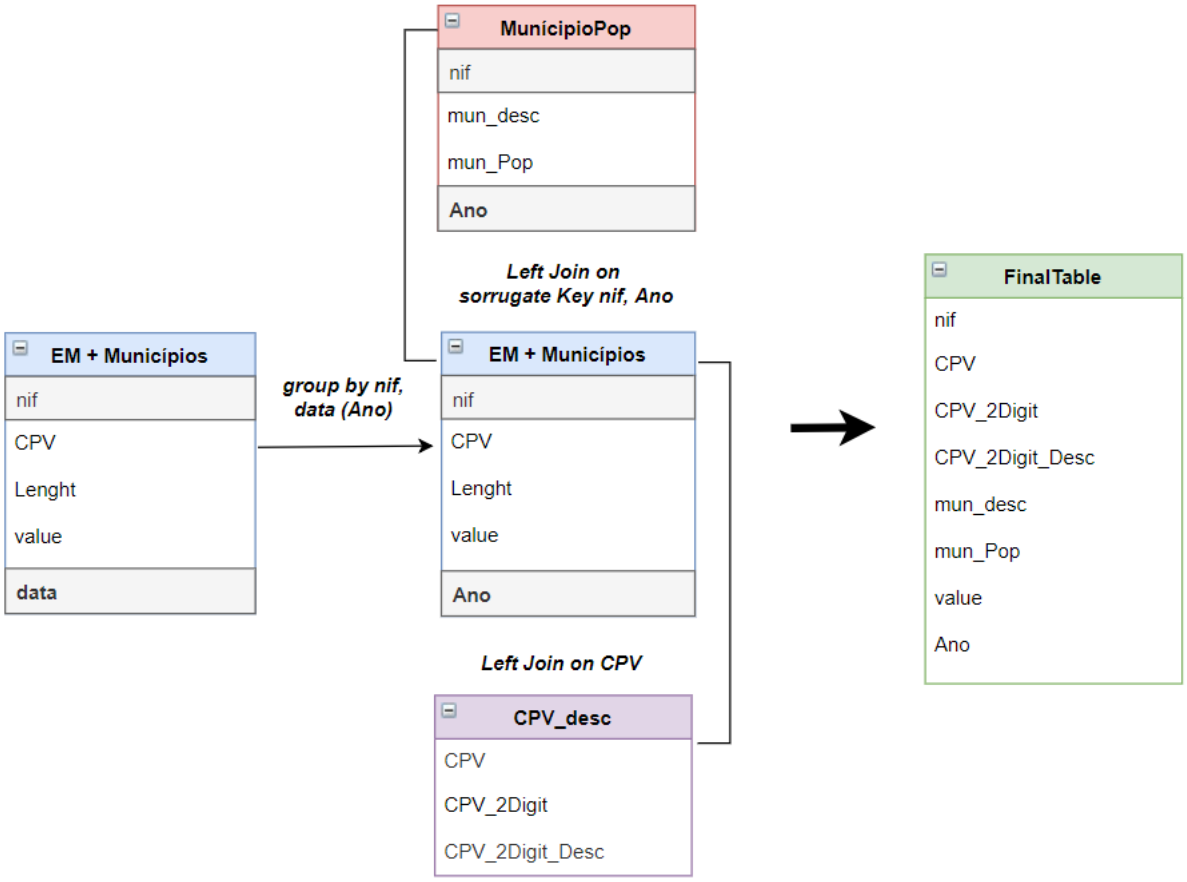


Figura 18 - *Merge* informação demográfica e contratual

Como referido na seção **1.1** os códigos CPV estão estruturados num código de **8 Dígitos** em que os 2 primeiros algarismos descrevem o primeiro nível da tipologia de despesa associada ao contrato (Divisões). Esta descrição da tipologia associada pode ser encontrada online no ficheiro em https://www.espap.pt/Documents/servicos/compras/CPV_2008.xls. Quantos mais algarismos forem considerados do respetivo código mais detalhe sobre a tipologia de despesa associada ao contrato será descrita.

Para este projeto foi considerado trabalhar com os dois primeiros dígitos, ou seja, o primeiro nível de tipos de despesa. Repare-se que já partimos de um número considerável de tipos de despesa tendo o primeiro nível **45** e o segundo **273** logo a dúvida seria sempre entre estes dois níveis. O

motivo principal desta decisão deve-se ao facto de apenas termos **88%** dos contratos classificados de nível 2 a 7 o que implicaria perder **12%** dos registos, correspondentes a **37.024** contratos, portanto iriamos ter uma grande perda de informação trabalhando com mais detalhe. O gráfico seguinte mostra o número de contratos por nível de CPV.

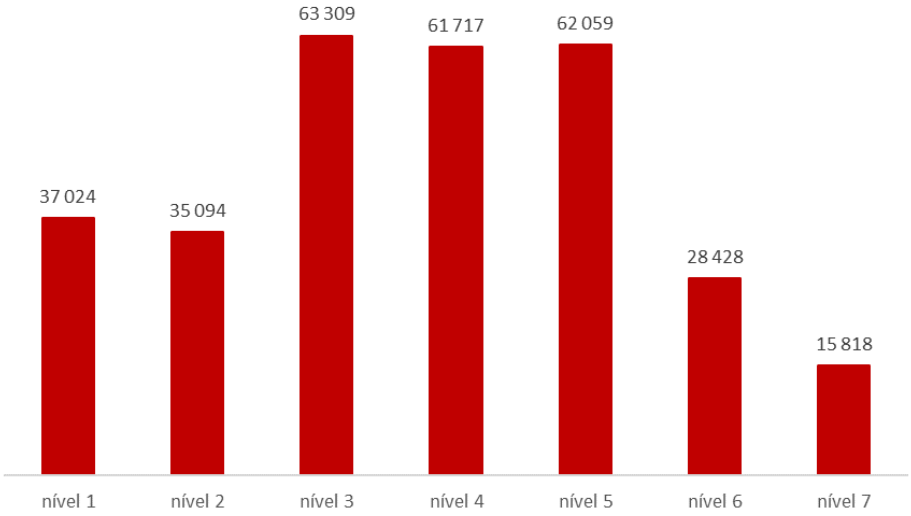
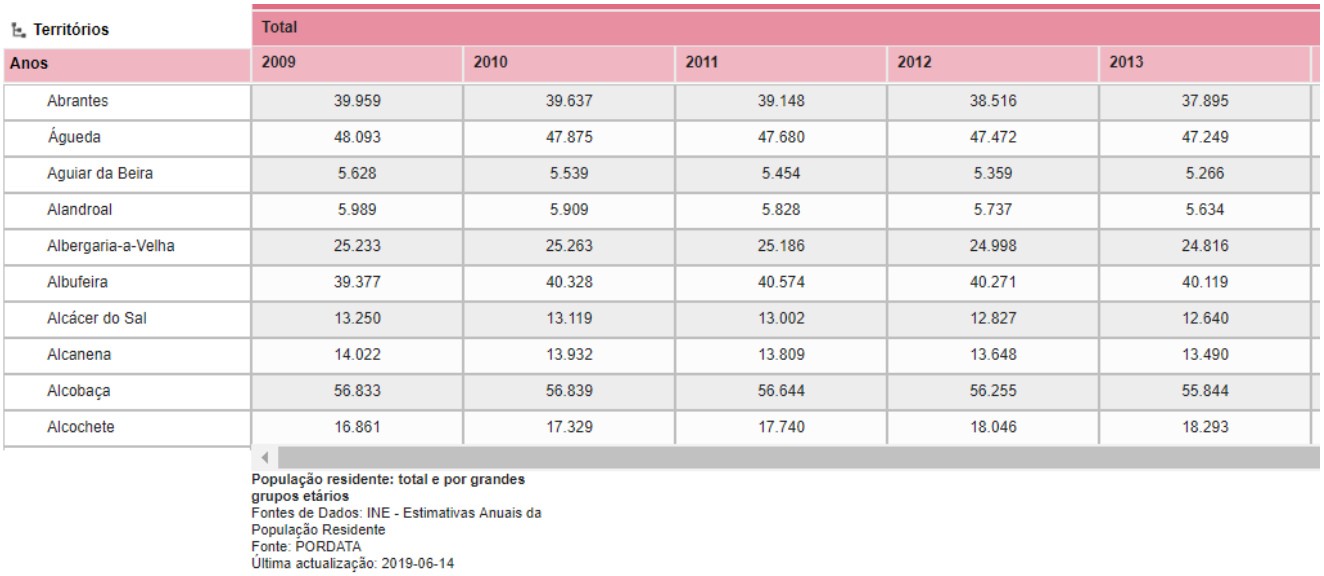


Figura 19 - Nº Contratos por nível de CPV

Por último foi necessário anexar informação demográfica. Sendo um dos objetivos deste trabalho perceber eventuais relações entre dimensão populacional dos municípios e despesa gasta pelos mesmos bem como evolução temporal tiveram de ser extraídos os dados relativos à população nos municípios ao ano. No site https://www.pordata.pt/DB/Municipios foi possível extrair esta informação (**Figura 19**). Esta informação foi extraída para um ficheiro *.xls.* Ainda foi necessária uma modificação desta tabela de forma a ficar apenas com apenas 4 colunas: *nif, Mun_Desc, Mun_Pop e Ano.*

Para podermos juntar a informação populacional tivemos de agregar a tabela de contratos ao nível ***nif*** e ***Ano*** de forma a podermos juntar as duas tabelas através da *sorrugate Key* ***'nif_Ano'*** como é mostrado na **Figura 20.**



Territórios	Total				
Anos	2009	2010	2011	2012	2013
Abrantes	39.959	39.637	39.148	38.516	37.895
Águeda	48.093	47.875	47.680	47.472	47.249
Aguiar da Beira	5.628	5.539	5.454	5.359	5.266
Alandroal	5.989	5.909	5.828	5.737	5.634
Albergaria-a-Velha	25.233	25.263	25.186	24.998	24.816
Albufeira	39.377	40.328	40.574	40.271	40.119
Alcácer do Sal	13.250	13.119	13.002	12.827	12.640
Alcanena	14.022	13.932	13.809	13.648	13.490
Alcobaça	56.833	56.839	56.644	56.255	55.844
Alcochete	16.861	17.329	17.740	18.046	18.293

População residente: total e por grandes grupos etários
Fontes de Dados: INE - Estimativas Anuais da População Residente
Fonte: PORDATA
Última actualização: 2019-06-14

Figura 20 - *PrintScreen* Site PORDATA: Excerto tabela relativo à dimensão populacional anual dos municípios em Portugal entre 2009 e 2013.

3.4.Análise Exploratória

Como foi apresentado na **Figura 18** as variáveis consideradas para análise têm a seguinte estrutura (**Tabela 6**):

Campo	**Descrição**	**Typ**
Nif	Nif Município	*Int*
CPV	CPV contrato	*Int*
CPV_2dig	Primeiros 2 algarismos do CPV	*Int*
CPV_2dig_des	Tipo de contrato associado aos 2 primeiros dígitos do	*Str*
Mun_desc	Descrição Município	*Str*
Mun_Pop	População Município	*Int*
Value	Valor do contrato	*Floa*
Ano	Ano do contrato	*Int*

Tabela 6 – Variáveis finais consideradas

De seguida são mostrados alguns mapas/gráficos onde é possível visualizar que municípios têm despesas mais elevadas bem como a evolução temporal entre 2009 e 2017. Por motivos de visualização optou-se por mostrar alguns gráficos a nível distrital.

As **Figura 21 e Figura 22** permitem visualizar que a maior parte da despesa é efetuada nas regiões de Lisboa, Porto e Algarve. Quando analisado ao Distrito podemos constatar que Lisboa e Porto juntos são 37% da despesa em Portugal.

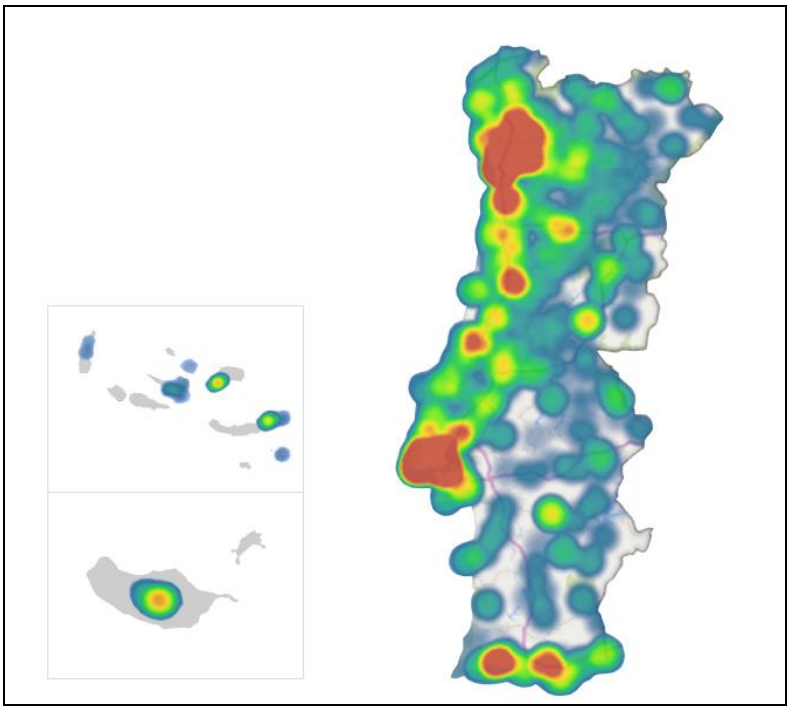


Figura 21 - *Heatmap* referente à despesa municipal

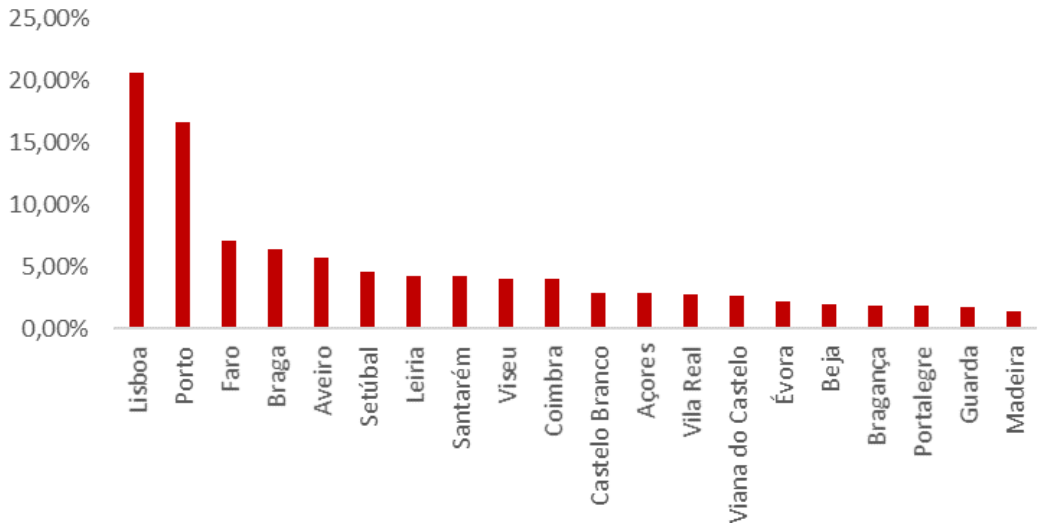


Figura 22 - Despesa agregada por Distrito

Na **Figura 23** é percetível que a despesa em todos os distritos teve uma quebra entre os anos de 2012 e 2015 devido ao efeito Troika sendo que o padrão é comum em todos os distritos.

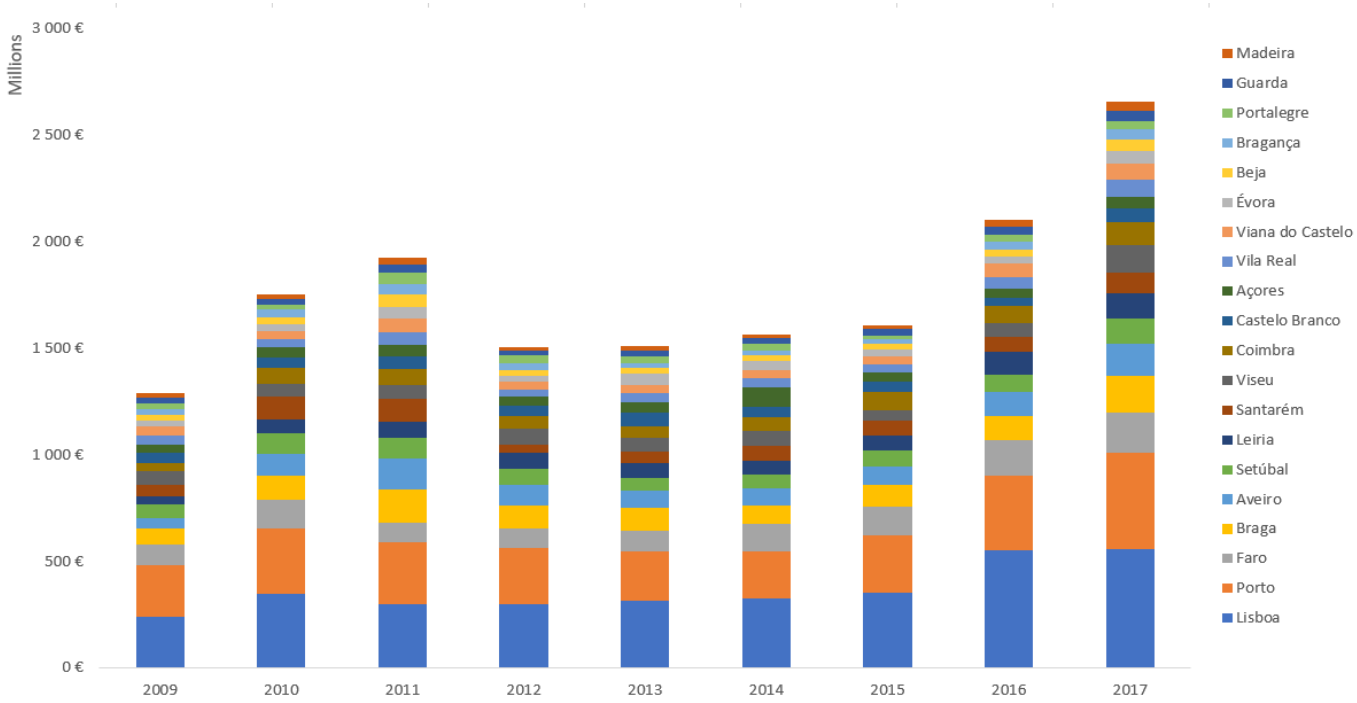


Figura 23 - Evolução anual da despesa a nível distrital

Por último na **Figura 24** é evidente a relação ***log-log*** entre despesa municipal e dimensão populacional semelhante ao longo dos anos entre os municípios associados. Ainda é possível visualizar que a mancha fica mais robusta ao longo do tempo sendo que alguns municípios inicialmente estão mais afastados da '*trend line'* acabam por aproximar nos últimos anos.

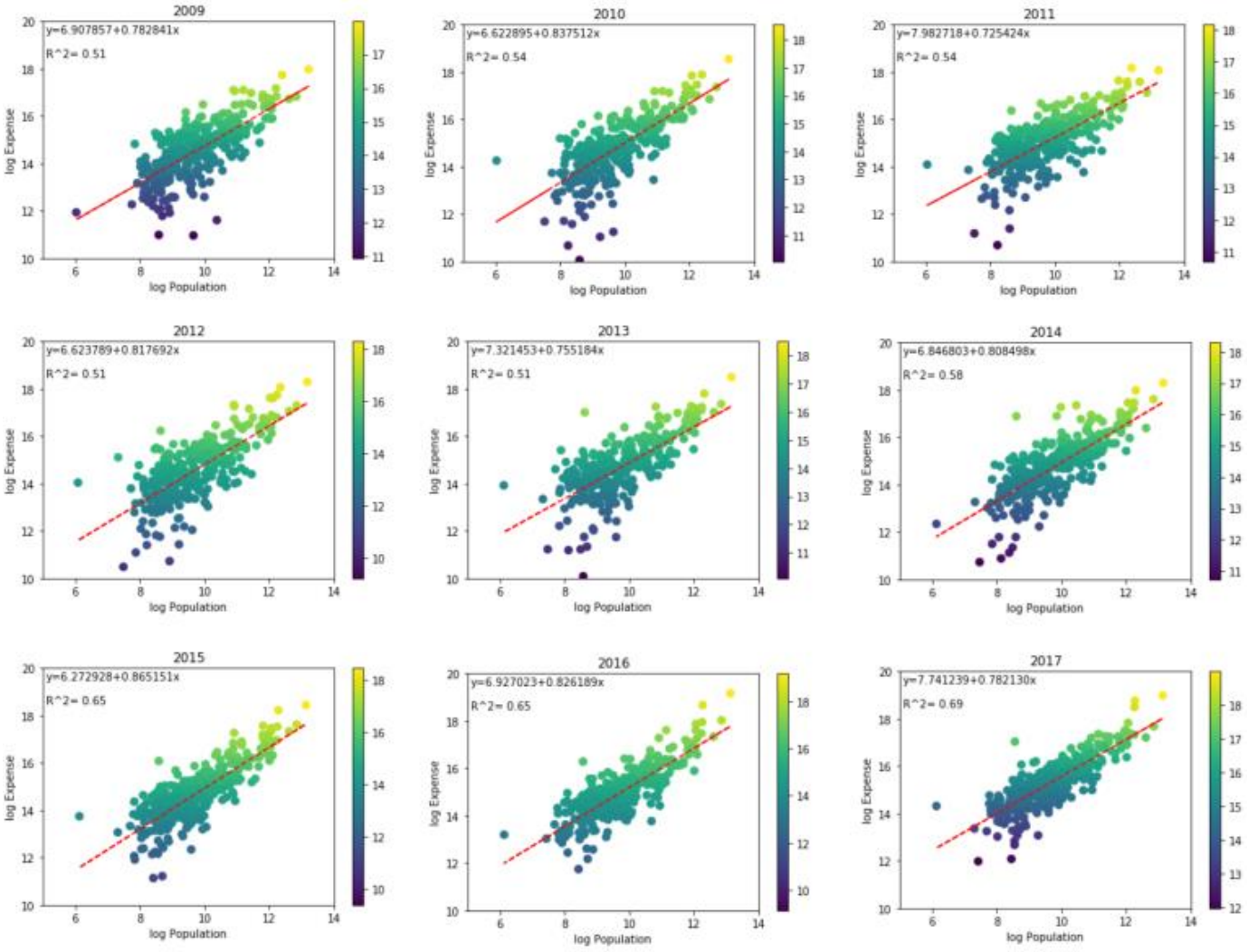


Figura 24 - Evolução temporal Despesa vs população municipal

TIPOS DE DESPESA

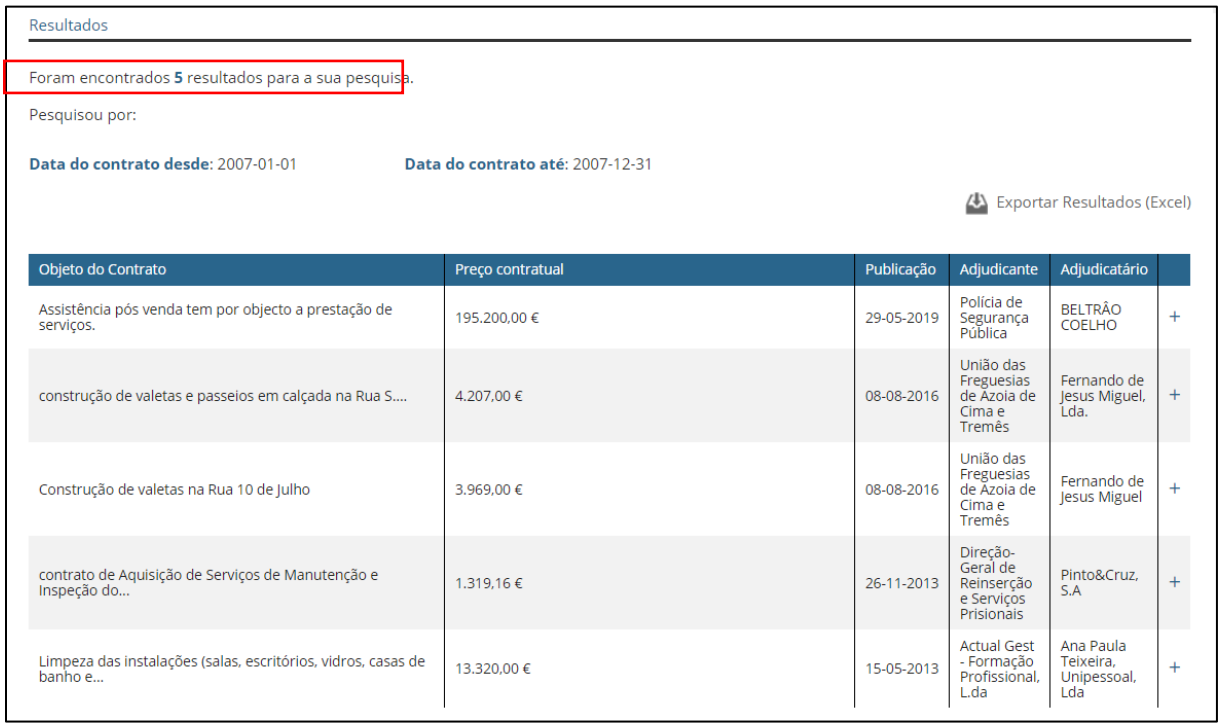
Tipo Despesa	€	%
Construção	8 163 682 032 €	51,3%
Produtos petrolíferos, combustíveis, electricidade e outras fontes de energia	859 655 990 €	5,4%
Serviços a empresas: direito, comercialização, consultoria, recrutamento, impressão e segurança	764 450 498 €	4,8%
Serviços relativos a águas residuais, resíduos, limpeza e ambiente	715 883 651 €	4,5%
Serviços de arquitectura, construção, engenharia e inspecção	636 814 235 €	4,0%
Serviços de hotelaria, restauração e comércio a retalho	464 092 414 €	2,9%
Equipamento e produtos auxiliares de transporte	386 898 423 €	2,4%
Serviços recreativos, culturais e desportivos	359 805 046 €	2,3%
Estruturas e materiais de construção; produtos auxiliares de construção (excepto aparelhos eléctricos)	333 755 321 €	2,1%
Serviços de reparação e manutenção	323 908 933 €	2,0%
Serviços de transporte (excl. transporte de resíduos)	301 747 026 €	1,9%
Serviços de TI: consultoria, desenvolvimento de software, Internet e apoio	250 858 169 €	1,6%
Serviços públicos	223 425 758 €	1,4%
Serviços de finanças e seguros	211 989 452 €	1,3%
Outros serviços comunitários, sociais e pessoais	192 217 834 €	1,2%
Mobiliário (incl. de escritório), acessórios, aparelhos domésticos (excl. iluminação) e produtos de limpeza	176 621 356 €	1,1%
Máquinas, equipamento e material de escritório e de informática, excepto mobiliário e pacotes de programas (software)	152 804 191 €	1,0%
Serviços de agricultura, silvicultura, horticultura, aquicultura e apicultura	135 253 071 €	0,9%
Pacotes de software e sistemas de informação	131 811 538 €	0,8%
Produtos alimentares, bebidas, tabaco e produtos afins	117 395 930 €	0,7%
Serviços de ensino e formação	95 644 484 €	0,6%
Serviços postais e de telecomunicações	88 922 704 €	0,6%
Maquinaria, aparelhagem, equipamento e consumíveis eléctricos; iluminação	78 065 184 €	0,5%
Equipamento de rádio, televisão, comunicação, telecomunicações e afins	70 246 615 €	0,4%
Equipamento laboratorial, óptico e de precisão (exc. óculos)	60 981 309 €	0,4%
Máquinas industriais	59 889 081 €	0,4%
Serviços relacionados com a administração pública, a defesa e a segurança social	58 560 365 €	0,4%
Material impresso e afins	55 192 107 €	0,3%
Exploração mineira, metais de base e produtos afins	42 591 064 €	0,3%
Instrumentos musicais, artigos de desporto, jogos, brinquedos, material para artesanato e actividades artísticas e acessórios	40 983 243 €	0,3%
Serviços de saúde e acção social	40 361 760 €	0,3%
Produtos químicos	38 297 771 €	0,2%
Serviços de instalação (excepto software)	37 622 677 €	0,2%
Maquinaria para extracção mineira e pedreiras, equipamento de construção	37 067 153 €	0,2%
Vestuário, calçado, malas e artigos de viagem, acessórios	35 460 995 €	0,2%
Produtos da agricultura, da pesca, da silvicultura e afins	28 012 382 €	0,2%
Serviços de investigação e desenvolvimento e serviços de consultoria conexos	25 545 916 €	0,2%
Serviços anexos e auxiliares dos transportes; serviços de agências de viagens	23 767 346 €	0,1%
Equipamento de segurança, combate a incêndios, polícia e defesa	18 471 295 €	0,1%
Serviços relacionados com as indústrias do gás e do petróleo	17 632 569 €	0,1%
Materiais têxteis, de couro, de plástico e de borracha	14 880 976 €	0,1%
Maquinaria agrícola	13 127 973 €	0,1%
Equipamento médico, medicamentos e produtos para cuidados pessoais	12 392 895 €	0,1%
Água captada e tratada.	6 706 633 €	0,0%
Serviços imobiliários	4 366 417 €	0,0%

Tabela 7 - Tipos de despesa

Como é possível ver na **Tabela 7,** o tipo de despesa 'Construção' ocupa o primeiro lugar com mais de 50% dos gastos municipais. O peso desta variável a nível municipal está em linha com a análise feita na introdução deste projeto sem filtrar ao nível Município.

3.5.Examinação de Outliers

Antes de entrar na análise de Outliers concretamente, é importante ter a noção da evolução do número de contratos lançados por ano. Este ponto é fundamental para perceber se estamos a lidar com *missing data*. Na **Figura 26** é possível verificar que há um aumento crescente do número de contratos lançados ao longo dos anos o que sugere que muitas despesas não terão sido lançadas em anos iniciais. Recorde-se que a publicação dos contratos online começou em 2007. Os anos de 2007 e 2008 não foram considerados para este projeto devido a tratarem-se de cerca 5 e 2457 respetivamente. Na **Figura 25** é possível ver um *snap shot* da página onde se encontram os contratos para o ano de 2007.



Resultados

Foram encontrados **5** resultados para a sua pesquisa.

Pesquisou por:

Data do contrato desde: 2007-01-01 **Data do contrato até**: 2007-12-31

Exportar Resultados (Excel)

Objeto do Contrato	Preço contratual	Publicação	Adjudicante	Adjudicatário	
Assistência pós venda tem por objecto a prestação de serviços.	195.200,00 €	29-05-2019	Polícia de Segurança Pública	BELTRÃO COELHO	+
construção de valetas e passeios em calçada na Rua S....	4.207,00 €	08-08-2016	União das Freguesias de Azoia de Cima e Tremês	Fernando de Jesus Miguel, Lda.	+
Construção de valetas na Rua 10 de Julho	3.969,00 €	08-08-2016	União das Freguesias de Azoia de Cima e Tremês	Fernando de Jesus Miguel	+
contrato de Aquisição de Serviços de Manutenção e Inspeção do...	1.319,16 €	26-11-2013	Direção-Geral de Reinserção e Serviços Prisionais	Pinto&Cruz, S.A	+
Limpeza das instalações (salas, escritórios, vidros, casas de banho e...	13.320,00 €	15-05-2013	Actual Gest - Formação Profissional, L.da	Ana Paula Teixeira, Unipessoal, Lda	+

Figura 25 - Nº de contratos disponíveis em 2007 no portal gov.pt

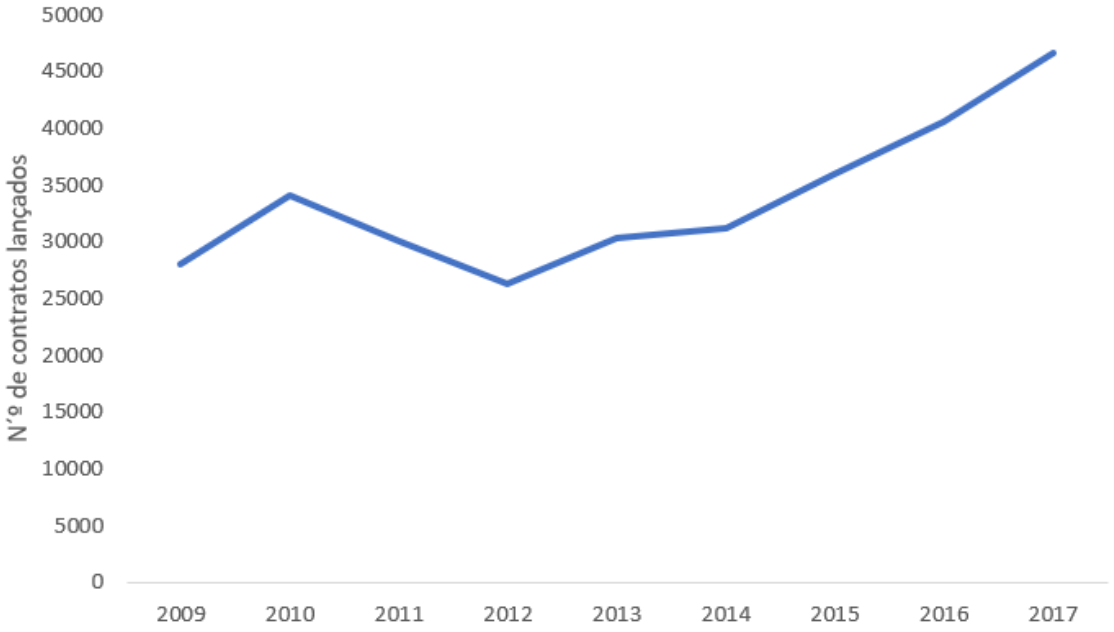


Figura 26 - Evolução anual do nº de contratos lançados no portal gov.pt relativos a Municípios e Empresas Municipais

Neste estudo apenas serão considerados os anos entre 2009 e 2017. Podemos ver que em 2009 claramente há um aumento significativo do número de contratos lançados face a 2008 que como mencionado no parágrafo anterior são 5 e 2457 respetivamente.

Começando por explorar a variável '*valor*' através de um histograma é possível visualizar que a distribuição é caracterizada por uma *positive skew*. Na **Figura 27** está representada um excerto do histograma por questões de visualização, do logaritmo da variável *'valor'*.

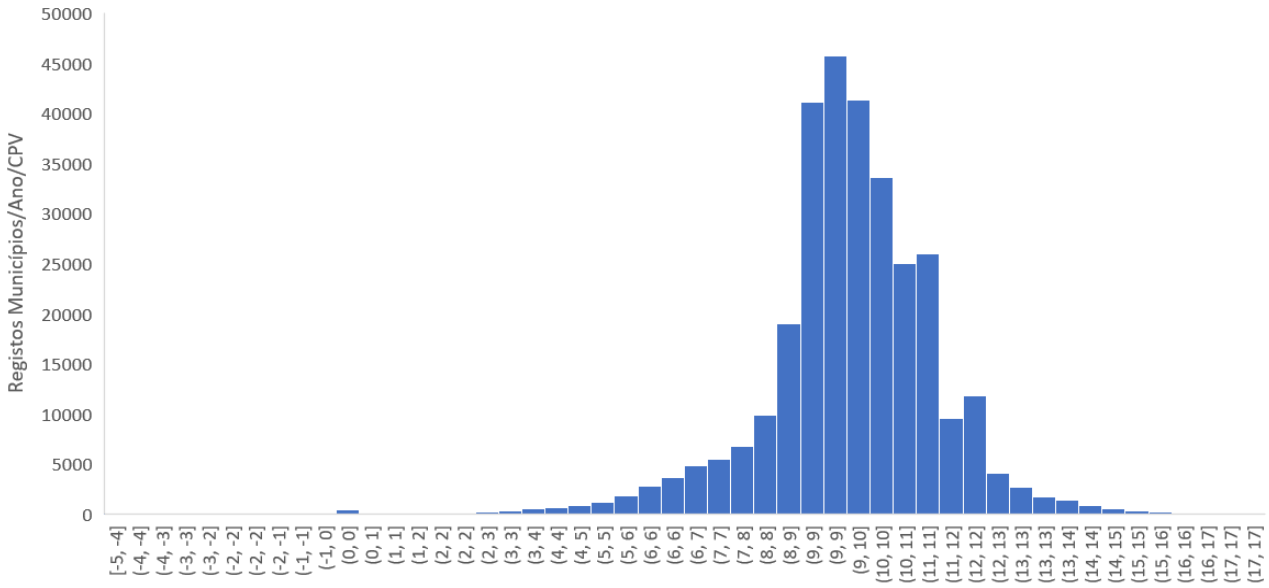


Figura 27 - Histograma logaritmo da despesa municipal 2009 – 2017

As **Figura 28**, **Figura 29** e **Figura 30** permitem observar a dispersão de despesa ao nível CPV-Ano para 3 dos 8 anos em análise, respetivamente 2009, 2014 e 2017. Para estes 3 anos é evidente a presença de pontos com valor fora do intervalo delimitado pelos **percentis 25% e 75%.** Este facto é evidente para a despesa referente ao **CPV 45: Construção.**

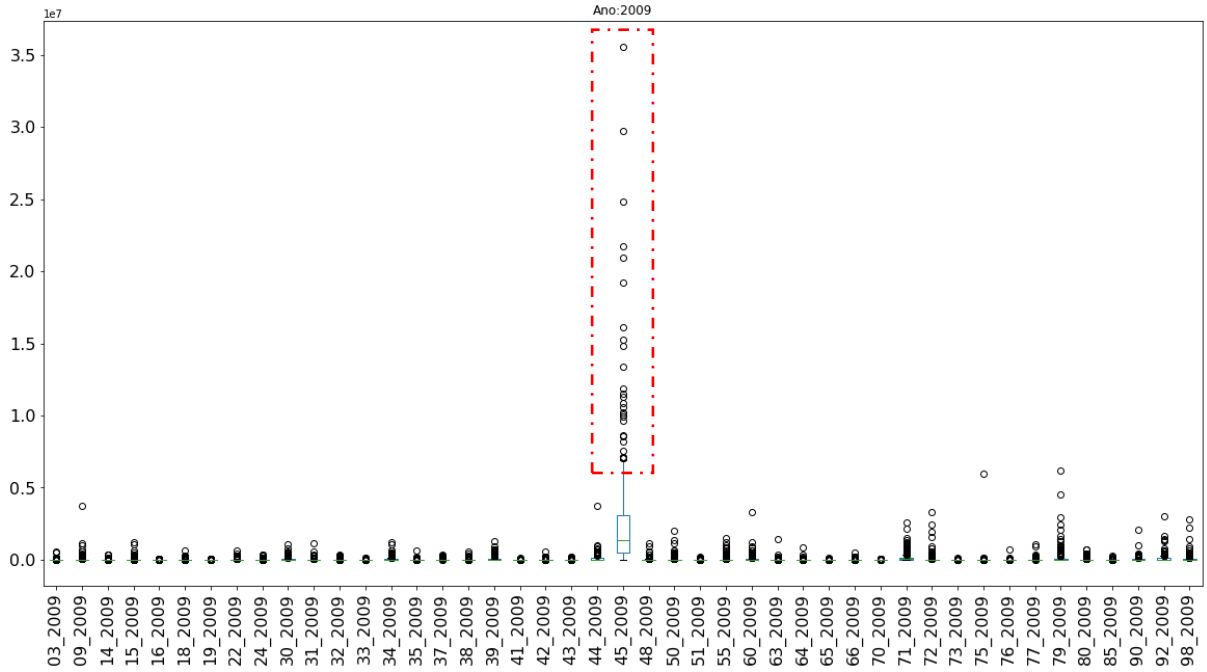


Figura 28 - *Box Plot* despesa municipal por CPV em 2009

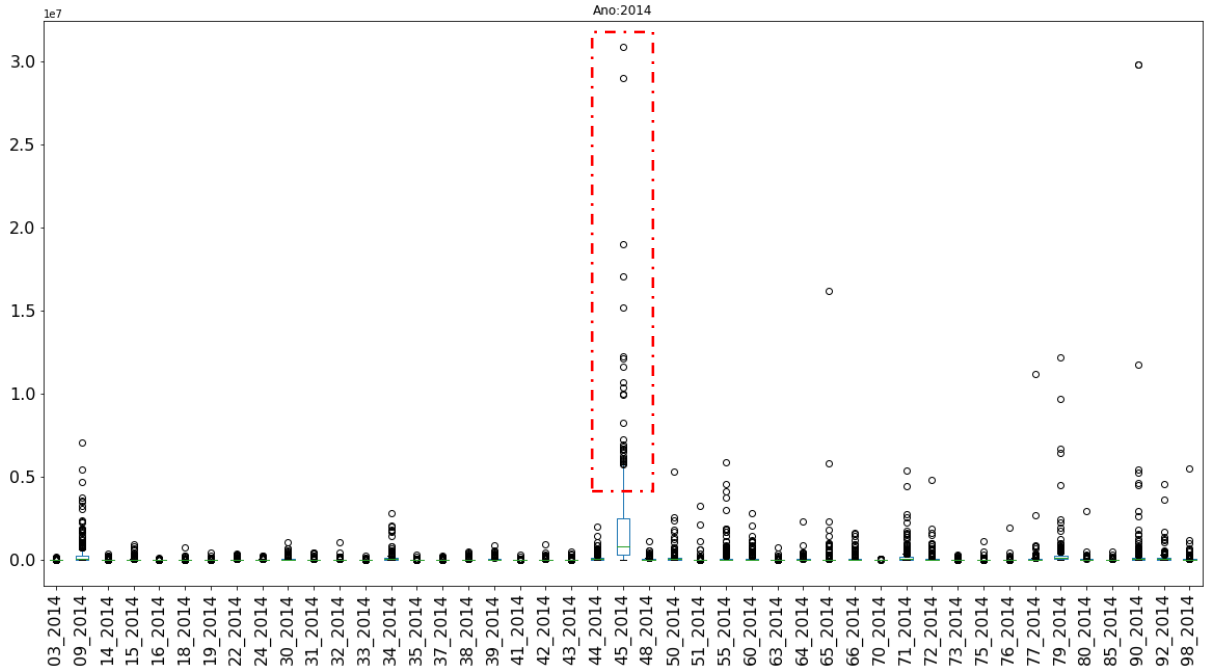


Figura 29 - *Box Plot* despesa municipal por CPV em 2014

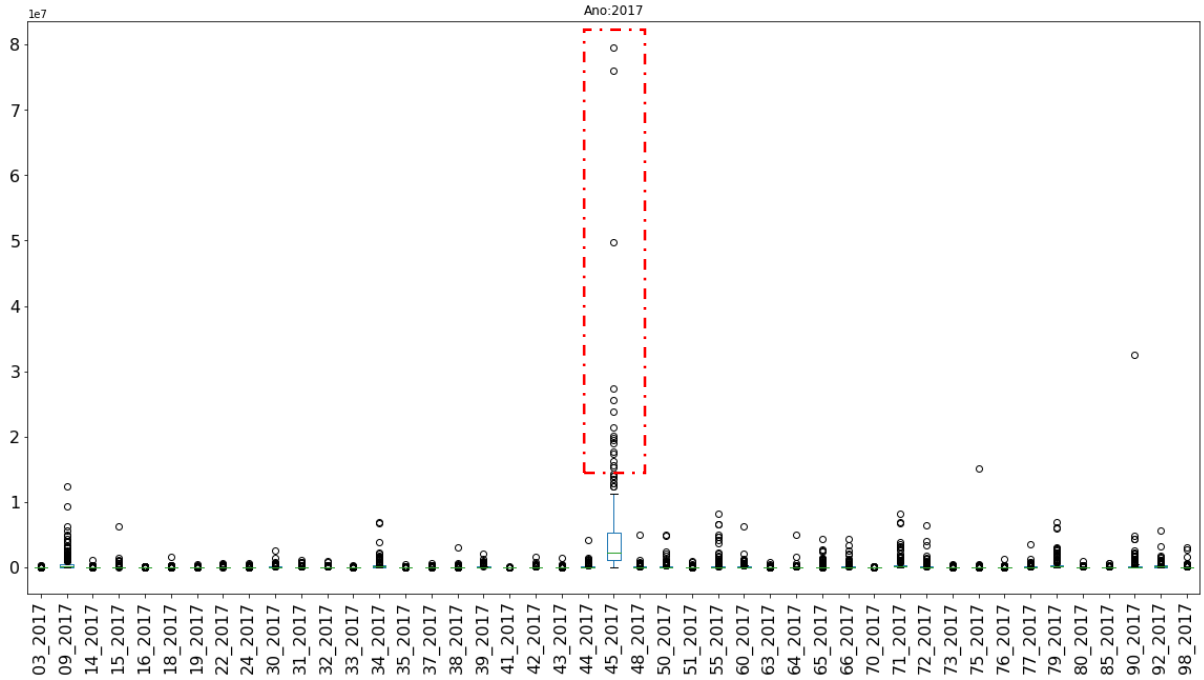


Figura 30 - *Box Plot* despesa municipal por CPV em 2017

Como já tínhamos visto na análise exploratória sem partir por tipo de contrato, a função que melhor explica a relação entre despesa total e dimensão da população é a multiplicativa $y = ax^b$ pelo que manteremos essa relação para os 45 tipos de despesa.

Para estudar a relação descrita no parágrafo anterior foi usado o modulo de *Python statsmodel.api* que contém inúmeras funções estatísticas entre elas o modelo ***Ordinary Least Squares (OLS)*** que será o modelo usado durante esta análise. As funções contruídas têm como objetivo devolver os coeficientes e métricas estatísticas relativas à relação ***'despesa'*** vs '***dimensão_população***'. A **Figura 30** ilustra as funções usadas com recurso à biblioteca ***OLS*** bem como uma descrição de cada uma.

```
def Fx_Regression (value, DimPop, Cpv_Year):
    X=sm.add_constant(DimPop)
    Regression = sm.OLS(value, X).fit()
    Regression_Coeff = Regression.params
    Regression_Coeff = Regression_Coeff.to_frame().T
    Regression_Coeff['Cpv_Year']=Cpv_Year
    Regression_Coeff.columns = ['const','Coeff', 'Cpv_Year']
    return Regression_Coeff

def Fx_Regression_Stats (value, DimPop, Cpv_Year):
    X=sm.add_constant(DimPop)
    Regression = sm.OLS(value, X).fit()
    Regression_Stats=Regression.pvalues
    Regression_Rsqr=Regression.rsquared
    Regression_Stats = Regression_Stats.to_frame().T
    Regression_Stats['Rsqr']=Regression_Rsqr
    Regression_Stats['Cpv_Year']=Cpv_Year
    Regression_Stats.columns = ['p-value_const','p-value_Coeff','Rsqr', 'Cpv_Year']
    return Regression_Stats


def Fx_Regression2(value, VarInd):
    X=sm.add_constant(VarInd)
    Regression = sm.OLS(value, X).fit()
    influence = Regression.get_influence()

    # model values
    model_fitted_y = Regression.fittedvalues

    studentized_residuals = influence.resid_studentized_external

    model_residuals = Regression.resid
    # normalized residuals
    model_norm_residuals = Regression.get_influence().resid_studentized_internal
    # absolute squared normalized residuals
    model_norm_residuals_abs_sqrt = np.sqrt(np.abs(model_norm_residuals))
    # absolute residuals
    model_abs_resid = np.abs(model_residuals)
    # leverage, from statsmodels internals
    model_leverage = Regression.get_influence().hat_matrix_diag
    # cook's distance, from statsmodels internals
    model_cooks = Regression.get_influence().cooks_distance[0]

    return [model_fitted_y, model_residuals, studentized_residuals,
            model_norm_residuals, model_norm_residuals_abs_sqrt,
            model_abs_resid, model_leverage, model_cooks]
```

Figura 31 – Funções estatísticas para estudo da relação *'**despesa'*** vs '***dimensão_população**'*.

Argumentos das funções ilustradas na **Figura 31**:

***Value* –** vetor relativo ao gasto monetário dos respetivos municípios ao nível CPV/Ano

***DimPop* –** vetor relativo à dimensão da população num determinado ano

***Cpv_Year* –** vetor de *string*s constantes com a identificação do ano e do tipo de despesa

Output das funções ilustradas na **Figura 31**:

FX_Regression - devolve um *array* com a constante e coeficientes da regressão entre Despesa e Dimensão da população. Ainda devolve o código do tipo de despesa concatenado com o ano.

FX_Regression_Stats – equivalente à **FX_Regression**, mas que devolve as estatísticas *p-value, coeficiente p-value,* r^2 e código da despesa concatenado com o ano.

FX_Regression2 – devolve as predições do modelo *log-log* e várias métricas relativas a resíduos.

De forma a melhorar a qualidade das relações *'**despesa'*** vs '***dimensão_população**'* foram eliminados todos os registos que geraram *'Studentized Residuals'* com valor fora do intervalo [-3,3] e registos com *Distância de Cook* superior a 1. Após a eliminação destes registos foram calculadas novas regressões. Este processo eliminou cerca de 382 registos, aproximadamente 0,7%. o que permitiu melhorar a qualidade das regressões ao nível de *R-squared.*

Na **Figura 32** é possível observar os valores *R-squared* antes e após remoção de Outliers para os anos 2011, 2013, 2015 e 2017.

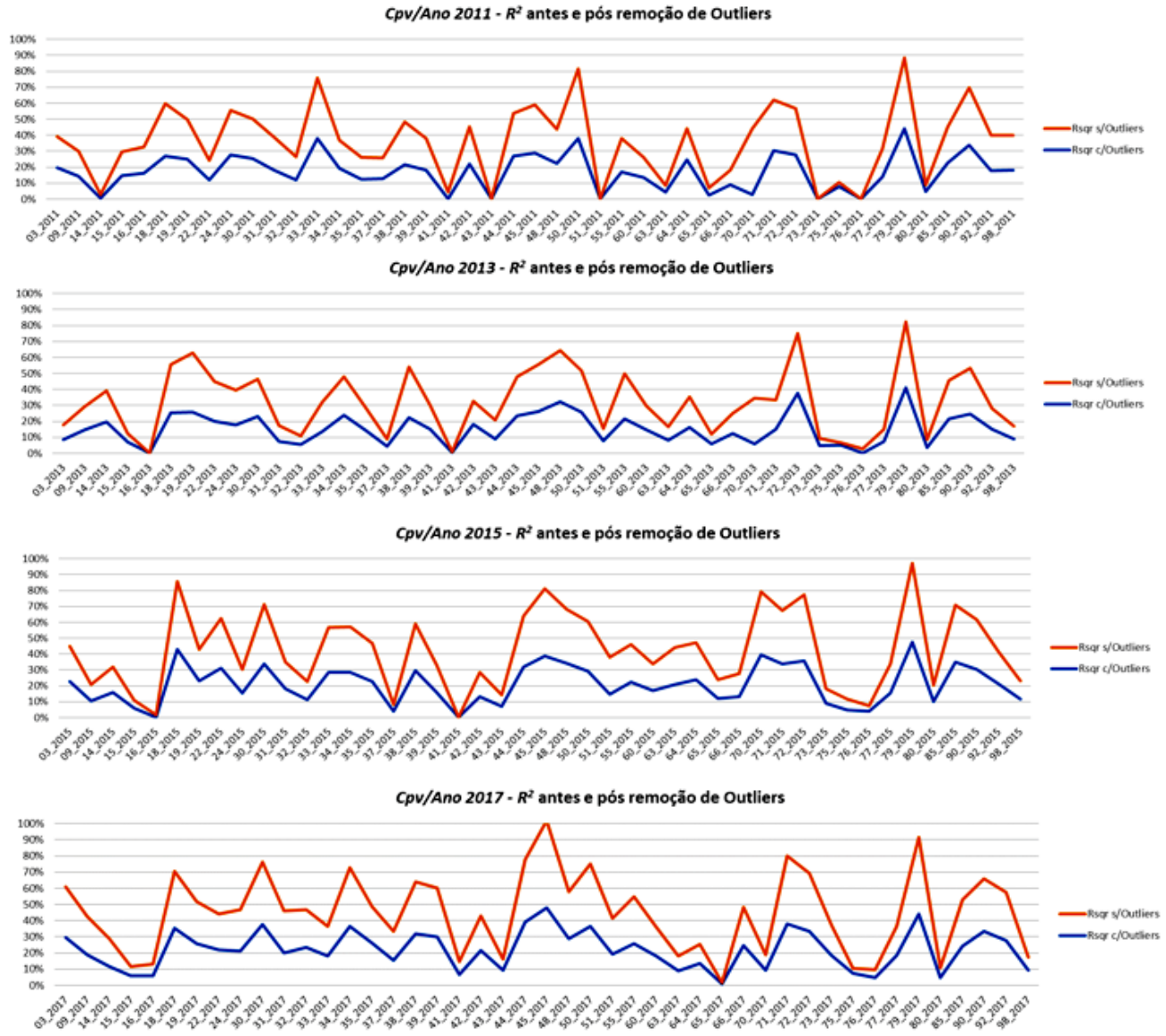


Figura 32 - *Rsqr* antes/após remoção de Outliers

3.6.Regressão Linear e Resíduos

Nesta fase são feitas as diversas regressões entre os diversos tipos de despesa e dimensão da população para cada ano logo estamos a falar **45*9** regressões, 45 tipos de contrato (nível 1 do CPV) multiplicado pelo número de anos, 9. Como referido no capítulo anterior, a função que melhor explica a relação entre despesa e dimensão da população é a multiplicativa $y = ax^b$ pelo que manteremos essa relação. De seguida pode-se visualizar o excerto de código com o qual foi executado e a respetiva explicação.

```
for i in Cpv_Year_List:
    Cpv=FinalTbl2.loc[FinalTbl2['Cpv_Year'] == i].reset_index()

    Cpv['log_value']=Cpv['value'].apply(np.log)
    Cpv['log_population']=Cpv['Population'].apply(np.log)
    Models=pd.concat([Models, Fx_Regression3(Cpv['log_value'], Cpv['log_population'], i)])
    Stats=pd.concat([Stats, Fx_Regression_Stats3(Cpv['log_value'], Cpv['log_population'], i)])
    Cpv=Cpv.merge(Models, how='left', on=['Cpv_Year'])
    Cpv=Cpv.merge(Stats, how='left', on=['Cpv_Year'])
    Cpv['model_fitted_y_new'], Cpv['model_residuals_new'],Cpv['studentized_residuals_new'], Cpv['model_norm_residuals_new'],
    Cpv['model_norm_residuals_abs_sqrt_new'], Cpv['model_abs_resid_new'],Cpv['model_leverage_new'], Cpv['model_cooks_new'] =
    Fx_Regression5(Cpv['log_value'], Cpv['log_population'])
    FinalTbl3=pd.concat([FinalTbl3, Cpv])

FinalTbl3['Prediction2']= FinalTbl3['const_y'].apply(np.exp)*(FinalTbl3['Population']**FinalTbl3['Coeff_y'])
FinalTbl3['Error2']=FinalTbl3['value']-FinalTbl3['Prediction2']
```

Figura 33 - Ciclo para cálculo das regressões *log-log* para cada CPV-Ano

Fazendo um breve resumo do código, o objetivo foi criar um ciclo pelo qual fossem sendo percorridos todos os elementos guardados numa lista única de todas as combinações **CPV_ANO** *(Cpv_Year_List).* Após cada iteração é aplicado a função *np.log* às variáveis *value* e *Population* ficando em condições de poder aplicar o modelo linear. Após estas duas transformações são invocadas as funções anteriormente descritas ***Fx_Regression***, ***Fx_Regression2 e Fx_Regression_Stats.*** A cada iteração vão sendo agregadas as linhas do output**.** Após finalizar o ciclo '*for'* é calculada a previsão real da despesa ($const * Population^{Coeff}$) e o respetivo resíduo em valor real.

3.7.Matriz de Distâncias

Tendo a variável necessária com a qual é construída a Matriz de distâncias: ***'Error'.*** Devido a existirem erros de diferentes magnitudes e sendo o objetivo deste projeto identificar grupos de CPV's com desvio semelhante houve a necessidade de normalizar esta variável ao nível do **CPV**. Este objetivo foi alcançado com a utilização da função ***RobustScaler*** disponível na biblioteca ***sklearn.preprocessing*.** A utilização deste *Scaler* deveu-se a ter maior capacidade no que toca a tratamento de Outliers. A **Figura 34** ilustra o processo.

```
ErrorTbl=pd.DataFrame(columns=['Mun_Year', 'Cpv_Year', 'CPV2', 'ErrorStd'])
Cpv_List=Cpv_Nif_value['CPV2'].unique().tolist()

for i in Cpv_List:

    Cpv=FinalTbl5.loc[FinalTbl5['CPV2'] == i].reset_index()      ## CPV
    Cpv['ErrorStd'] = RobustScaler().fit_transform(Cpv[['Error3']])
    Cpv = Cpv[['Mun_Year', 'Cpv_Year', 'CPV2','ErrorStd']]
    ErrorTbl=pd.concat([ErrorTbl, Cpv])
```

Figura 34 - Normalização dos erros ao nível CPV

Após a variável *'Error'* estar normalizada ao nível do CPV foi necessário rearranjar o *dataset* de forma a ficarmos com a coluna *CPV_Ano* nas colunas. Para isto foi usado a função *pivot_table* do modulo ***Pandas*** disponível em *Python*. Desta forma ficamos com uma tabela de $\mathbf{45 * 2766}$ isto é, **45** CPV's e $\mathbf{9 * 308}$ Municípios-Ano.

Devido ao facto de existirem muitos municípios que não têm despesa para um determinado ano num CPV não seria correto calcular a distância entre 2 CPV's diretamente entre os vetores de 2766 coordenadas em que muitos seriam '*Nan'*. A solução foi calcular individualmente as distâncias para todos os pares de CPV's, ou seja, ${}^{45}C_2$. O Objetivo foi poder remover todos os registos *'Município-Ano'* que não teriam despesa num ou nos dois CPV's e assim aproveitar o máximo de registos com valor superior a zero em cada combinação de contratos para calcular a respetiva distância. Para este efeito foi utilizada a biblioteca *itertools* disponível em *Python* que contem a função *combinations* que retorna uma lista com todas as combinações de um conjunto. Após a remoção dos registos com valor a zero é calculada a distância **Euclidiana ajustada** entre ambos os CPV's. De seguida esta tabela de distâncias é transformada na respetiva Matriz de Distâncias final *45x45* através da função *squareForm* disponível na biblioteca *scipy.spacial.*

```
FinalDistances=pd.DataFrame(columns=['Cpv1', 'Cpv2', 'dist'])
cc = list(combinations(DistanceTbl.columns,2))
for c in cc:

    temp=DistanceTbl.reset_index(level=0)
    CpvTbl=temp[['Mun_Year', c[0], c[1]]]
    CpvTbl = CpvTbl[CpvTbl[c[0]] != 0]
    CpvTbl = CpvTbl[CpvTbl[c[1]] != 0]
    dist = np.linalg.norm(CpvTbl[c[0]]- CpvTbl[c[1]])*(math.sqrt(2766/len(CpvTbl.index)))
    CpvTuple=pd.DataFrame(list(c)).transpose()
    CpvTuple.columns = ['Cpv1', 'Cpv2']
    distDf = pd.DataFrame(columns=['dist'])
    distDf.loc[0] = dist
    CpvTuple.join(distDf)
    FinalDistances=pd.concat([FinalDistances, CpvTuple.join(distDf)])

distanceMatrix=np.array(FinalDistances['dist'])
distanceMatrix_SF=scipy.spatial.distance.squareform(distanceMatrix, force='no', checks=True)
```

Figura 35 - Código para gerar matriz de Distâncias

Podemos observar um excerto (**Tabela 8**) da matriz de correlações apenas com os primeiros 10 CPV's gerada pelo código em cima. Esta matriz tem uma dimensão de **45x45**.

	0	1	2	3	4	5	6	7	8	9	10
0	0	326.848	131.576	248.09	118.228	307.177	164.478	176.246	243.383	159.367	219.054
1	326.848	0	280.771	371.507	342.961	351.599	435.474	282.675	349.319	245.202	288.459
2	131.576	280.771	0	194.303	138.794	245.468	185.206	164.508	215.845	134.267	167.106
3	248.09	371.507	194.303	0	149.012	391.228	297.806	277.163	284.241	247.921	304.186
4	118.228	342.961	138.794	149.012	0	391.224	146.509	170.155	252.569	184.6	268.738
5	307.177	351.599	245.468	391.228	391.224	0	356.572	290.32	316.094	184.13	242.068
6	164.478	435.474	185.206	297.806	146.509	356.572	0	209.109	209.185	209.755	271.142
7	176.246	282.675	164.508	277.163	170.155	290.32	209.109	0	253.657	163.542	219.555
8	243.383	349.319	215.845	284.241	252.569	316.094	209.185	253.657	0	202.916	275.269
9	159.367	245.202	134.267	247.921	184.6	184.13	209.755	163.542	202.916	0	152.947
10	219.054	288.459	167.106	304.186	268.738	242.068	271.142	219.555	275.269	152.947	0

Tabela 8 - Excerto da matriz de distâncias entre CPV's

A figura**Figura 36** - *Heatmap* de Matriz de Distâncias seguinte (**Figura 36**) permite começar a visualizar de forma mais clara que CPV's estão mais relacionados. Esta visualização foi conseguida através da função *heatmap* onde os quadrados azuis escuros significam maior distância entre os CPV's. É possível constatar que alguns CPV's claramente se encontram afastados dos restantes (Setas vermelhas).

Nota para o facto de agora a referência aos CPV's ser feita através do ID entre 0 e 44 e não dos originais. Os mais evidentes são os CPV's com **ID 26, 30 e 42** que representam respetivamente:

- **Serviços de hotelaria, restauração e comércio a retalho**
- **Serviços Públicos.**
- **Serviços relativos a águas residuais, resíduos, limpeza e ambiente**

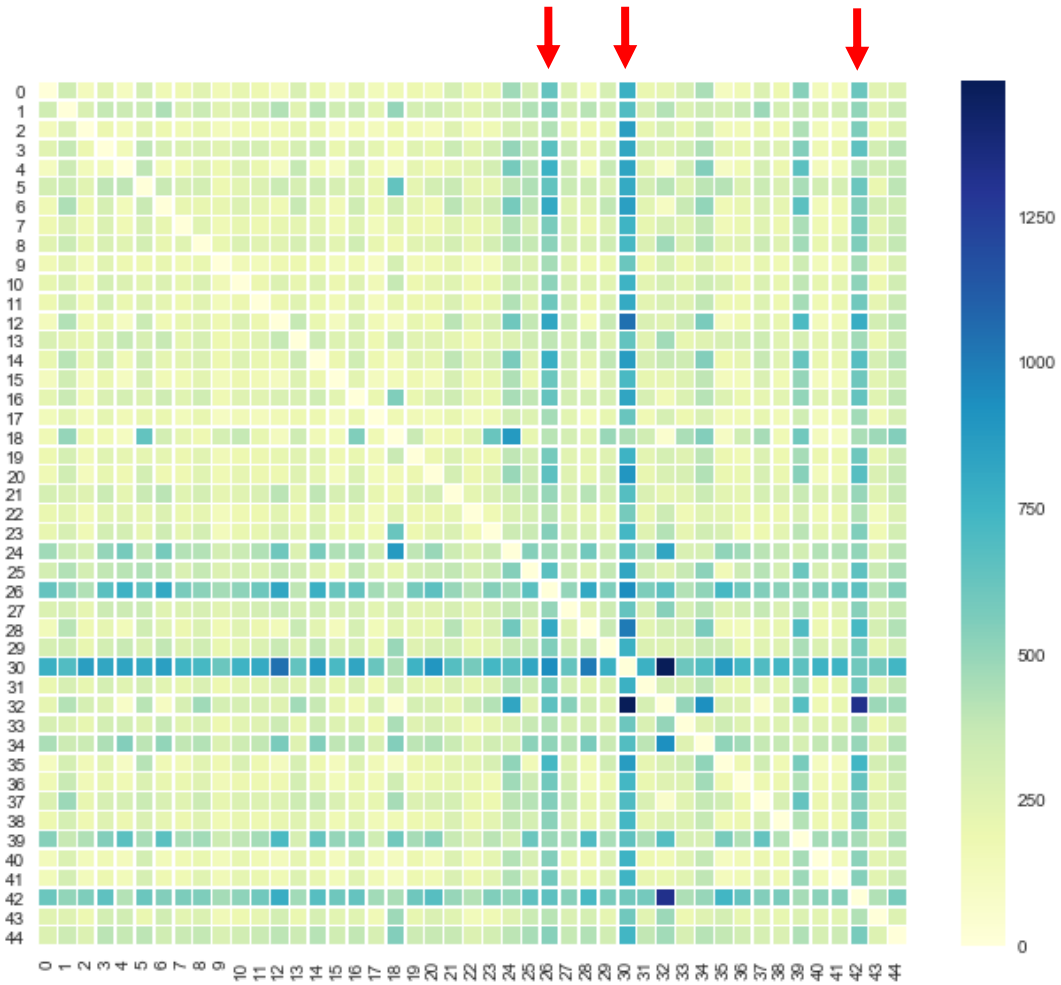


Figura 36 - *Heatmap* de Matriz de Distâncias e identificação dos CPV's mais distantes

3.8.Aplicação de algoritmos de Clustering

No seguimento do capítulo anterior, é nesta fase que iram ser testados três algoritmos distintos para segmentar os CPV's. Para tal, irá ser usada a matriz de distâncias anteriormente descrita. Devido a estarmos a trabalhar diretamente com a matriz de distâncias houve a limitação no uso de alguns algoritmos, como por exemplo o *K-Means*, que apenas trabalham com o *dataset* no formato de registos nas linhas e variáveis nas colunas. Os algoritmos que irão ser testados são:

1. *Hierárquico*
2. *K-Medoids*
3. *K-Means* após *MultiDimensionalScaling*

Na seção **4.1. Comparação e validação de algoritmos** será feita a avaliação e comparação dos algoritmos testados bem como a qualidade dos segmentos. Este processo será feito com recurso ao método ***Silhouette Score.***

ALGORITMO HIERÁRQUICO

Este algoritmo foi o primeiro a ser testado. Para obter os respetivos *Clusters* e dendrogramas foi usada a função *linkage* e *dend* presentes na biblioteca *scipy.Cluster.hierarchy*. o método usado foi o *Ward* cuja característica principal é minimizar a variância.

```
Z = linkage(distanceMatrix, 'ward')
f, ax = plt.subplots(figsize=(10, 15))
dend = dendrogram(Z, color_threshold=800, leaf_font_size=14, orientation='right', labels = CPV_RECODED2.index)
cutOff = sch.fcluster(Z, 1100, 'distance')
cluster_output = pd.DataFrame({'CPV':CPV_RECODED2.CPV , 'cluster':cutOff})
```

Figura 37 - Código Algoritmo Hierárquico

As **Figura 38** e **Figura 39** mostram os dendrogramas com divisão para 2 e 3 *Clusters*. A avaliação sobre o número de *Clusters* e avaliação da qualidade será feita na seção seguinte.

Figura 38 - Dendrogama 2 *Clusters*

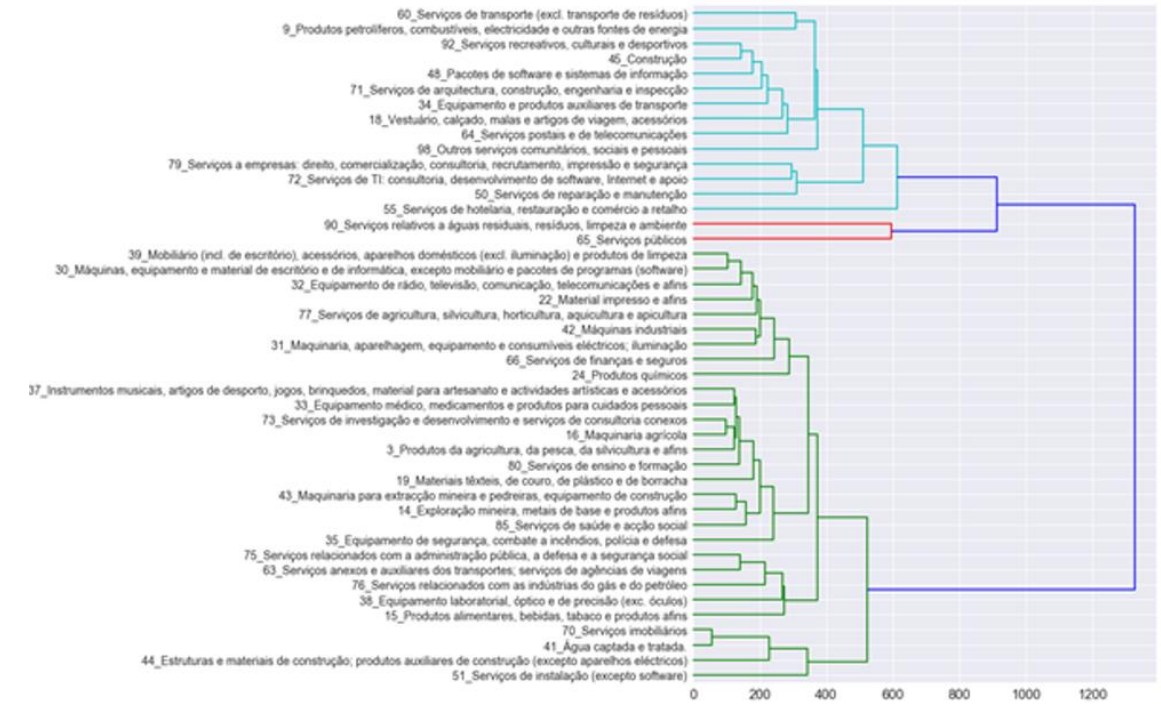


Figura 39 - Dendrograma 3 *Clusters*

ALGORITMO *K-MEDOIDS*

Este algoritmo teve de ser testado várias vezes devido à possibilidade de produzir *Clusters* diferentes a cada iniciação devido aos *Medoids* iniciais serem *random*.

Foi utilizada a biblioteca *pyClustering* e a função *K-Medoids.* O algoritmo foi executado 3 vezes para 2 e 3 *Clusters* com *random Medoids* iniciais.

```
initial_medoids = rd.sample([l for l in range(len(distanceMatrix_SF))], 3)
kmedoids_instance = kmedoids(KmedoidsMatrix, initial_medoids, data_type='distance_matrix')
# Run cluster analysis.
kmedoids_instance.process()
medoids = kmedoids_instance.get_medoids()
clusters = kmedoids_instance.get_clusters()
```

Figura 40 - Código *K-Medoids*

A **Tabela 9** mostra a distribuição dos CPV's nos respetivos *Clusters:*

ID	CPV	2 Clusters			3 Clusters		
		Run1	Run2	Run3	Run1	Run2	Run3
0	3	2	1	1	3	3	3
1	9	1	1	1	3	3	3
2	14	2	1	1	3	3	3
3	15	2	1	2	3	3	2
4	16	2	1	2	1	3	2
5	18	2	2	1	3	2	1
6	19	2	1	1	1	3	3
7	22	2	1	1	3	3	3
8	24	2	1	2	3	3	2
9	30	2	1	1	3	3	3
10	31	2	1	1	3	3	3
11	32	2	1	1	3	3	3
12	33	2	1	2	3	3	2
13	34	1	1	1	3	3	3
14	35	2	1	2	3	3	2
15	37	2	1	1	3	3	3
16	38	2	1	1	1	3	3
17	39	2	1	1	3	3	3
18	41	2	1	2	1	3	2
19	42	2	1	1	3	3	3
20	43	2	1	1	3	3	3
21	44	2	1	2	3	1	2
22	45	1	1	1	3	3	3
23	48	2	1	1	3	3	3
24	50	1	2	1	2	3	3
25	51	2	1	2	1	3	2
26	55	1	2	2	3	3	2
27	60	2	1	1	3	3	3
28	63	2	1	1	3	3	3
29	64	2	1	1	3	3	3
30	65	1	2	2	3	3	2
31	66	2	1	1	3	3	3
32	70	2	1	2	1	1	2
33	71	1	2	1	3	3	3
34	72	1	2	1	3	3	3
35	73	2	1	2	3	3	2
36	75	2	1	1	3	3	3
37	76	1	1	1	1	3	3
38	77	2	1	1	3	3	3
39	79	1	2	1	2	3	3
40	80	2	1	2	3	3	2
41	85	2	1	2	3	3	2
42	90	1	2	2	2	3	2
43	92	1	2	1	3	3	3
44	98	1	2	1	3	3	3

Tabela 9 - *Clusters K-Medoids*

MDS & K-MEANS

Devido ao facto de estarmos a trabalhar com uma matriz de distâncias especial que excluiu os missing values para o cálculo de todas as combinações de CPV's não foi possível usar o algoritmo *K-Means* diretamente sobre o *dataset* de **45x2766**. Uma maneira de contornar esta situação foi a utilização do algoritmo *Multidimensional Scaling* que permite reduzir a dimensionalidade neste caso de 2766 para 2 dimensões mantendo as distâncias originais. A **Figura 41** e **Figura 42** ilustram o código e o o *plot* dos CPV's nas novas coordenadas geradas pelo *MDS* respetivamente.

```
mds = manifold.MDS(n_components=2, dissimilarity="precomputed", random_state=3)
mdsResult = mds.fit(distanceMatrix_SF)

coords = mdsResult.embedding_

Cpvs=CPV_RECODED['CPV'].tolist()

plt.subplots(figsize=(20, 15))
plt.subplots_adjust(bottom = 0.1)
plt.scatter(coords[:, 0], coords[:, 1], marker = 'o')
#

for label, x, y in zip(Cpvs, coords[:, 0], coords[:, 1]):
    plt.annotate(
        label,
        xy = (x, y), xytext = (-20, 20),
        textcoords = 'offset points', ha = 'right', va = 'bottom',
        bbox = dict(boxstyle = 'round,pad=0.5', fc = 'yellow', alpha = 0.5),
        arrowprops = dict(arrowstyle = '->', connectionstyle = 'arc3,rad=0'))

plt.show()
```

Figura 41 – Código algoritmo *MDS*

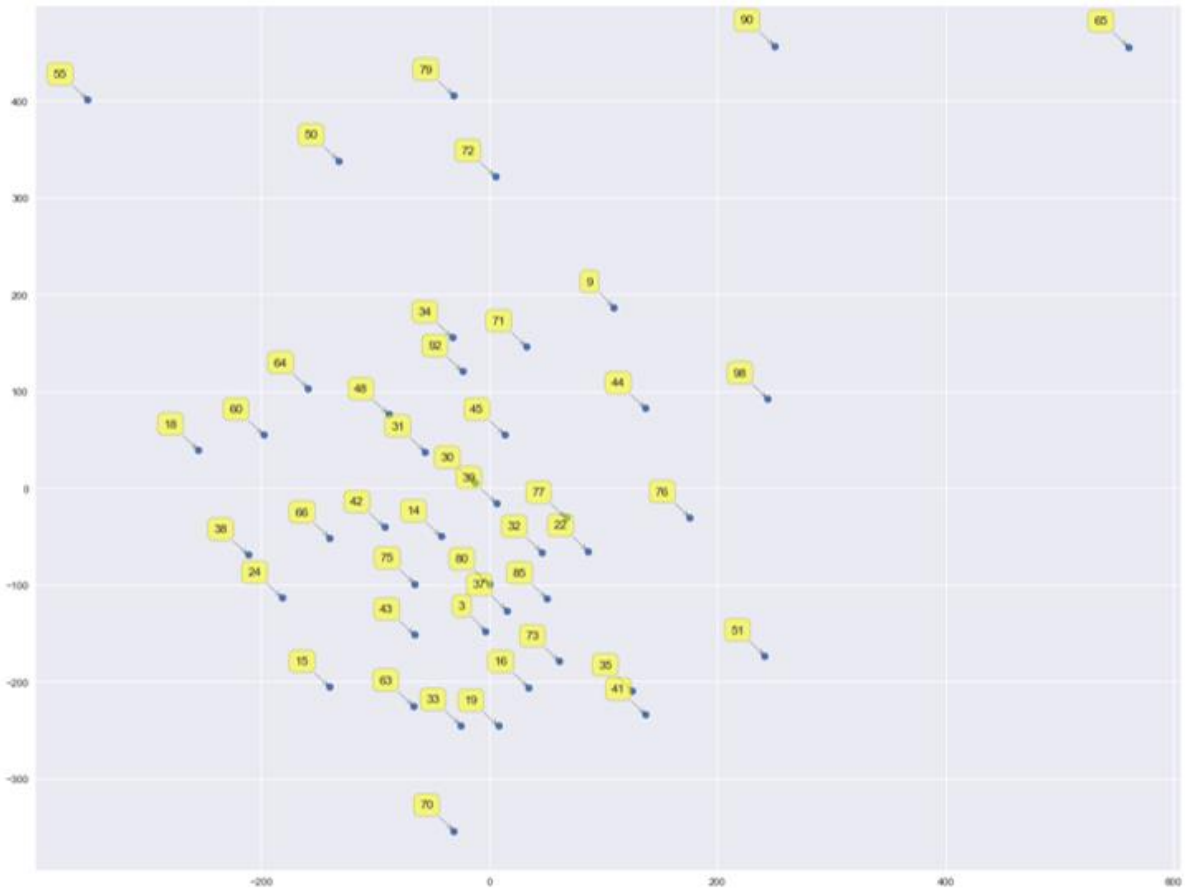


Figura 42 - Visualização CPV's após aplicação do *MDS* para duas dimensões

Após este processo testamos o algoritmo *K-Means* já sobre o *dataset* de duas dimensões gerado pelo *MDS*. Como é possível observar na **Figura 42**, o número de *Clusters* aponta para 2 ou 3 no máximo. Do mesmo modo que foi executado o algoritmo *K-Medoids* foram executadas 3 *runs* para dois e três *Clusters* com *random seeds.*

```
kmeans = KMeans(n_clusters = 2, init = 'random')
kmeans.fit(X)
wcss.append(kmeans.inertia_)

clusters=kmeans.labels_
clusters_number=np.unique(clusters)
clusters_number=np.size(clusters_number)
clust_number.append(clusters_number)
```

Figura 43 - Código Algoritmo *K-Means*

A **Tabela 10** mostra a distribuição dos CPV's nos respetivos *Clusters:*

ID	CPV	2 Clusters			3 Clusters		
		Run1	Run2	Run3	Run1	Run2	Run3
0	3	1	1	1	0	1	1
1	9	0	0	0	2	2	2
2	14	1	1	1	0	1	1
3	15	1	1	1	0	1	1
4	16	1	1	1	0	1	1
5	18	1	1	0	2	0	0
6	19	1	1	1	0	1	1
7	22	1	1	1	0	1	1
8	24	1	1	1	0	1	0
9	30	1	1	1	0	1	0
10	31	1	1	0	2	0	0
11	32	1	1	1	0	1	1
12	33	1	1	1	0	1	1
13	34	0	0	0	2	0	0
14	35	1	1	1	0	1	1
15	37	1	1	1	0	1	1
16	38	1	1	1	0	1	0
17	39	1	1	1	0	1	1
18	41	1	1	1	0	1	1
19	42	1	1	1	0	1	0
20	43	1	1	1	0	1	1
21	44	0	0	0	2	2	2
22	45	1	1	0	2	0	0
23	48	1	1	0	2	0	0
24	50	0	0	0	2	0	0
25	51	1	1	1	0	1	1
26	55	0	0	0	2	0	0
27	60	1	1	0	2	0	0
28	63	1	1	1	0	1	1
29	64	1	1	0	2	0	0
30	65	0	0	0	1	2	2
31	66	1	1	1	0	1	0
32	70	1	1	1	0	1	1
33	71	0	0	0	2	0	0
34	72	0	0	0	2	0	2
35	73	1	1	1	0	1	1
36	75	1	1	1	0	1	1
37	76	1	1	1	0	1	1
38	77	1	1	1	0	1	1
39	79	0	0	0	2	0	2
40	80	1	1	1	0	1	1
41	85	1	1	1	0	1	1
42	90	0	0	0	1	2	2
43	92	0	0	0	2	0	0
44	98	0	0	0	2	2	2

Tabela 10 - *Clusters* ***K-Means***

4.Discussão de Resultados

4.1.Comparação e validação de algoritmos

Nesta seção é discutida a qualidade da segmentação produzida pelos algoritmos mencionados no capítulo anterior. Na tabela seguinte podemos fazer uma comparação dos diversos algoritmos e como foram assignados os CPV's aos respetivos *Clusters*. Na **Figura 44** podemos visualizar o *Silhouete Score* para cada segmentação.

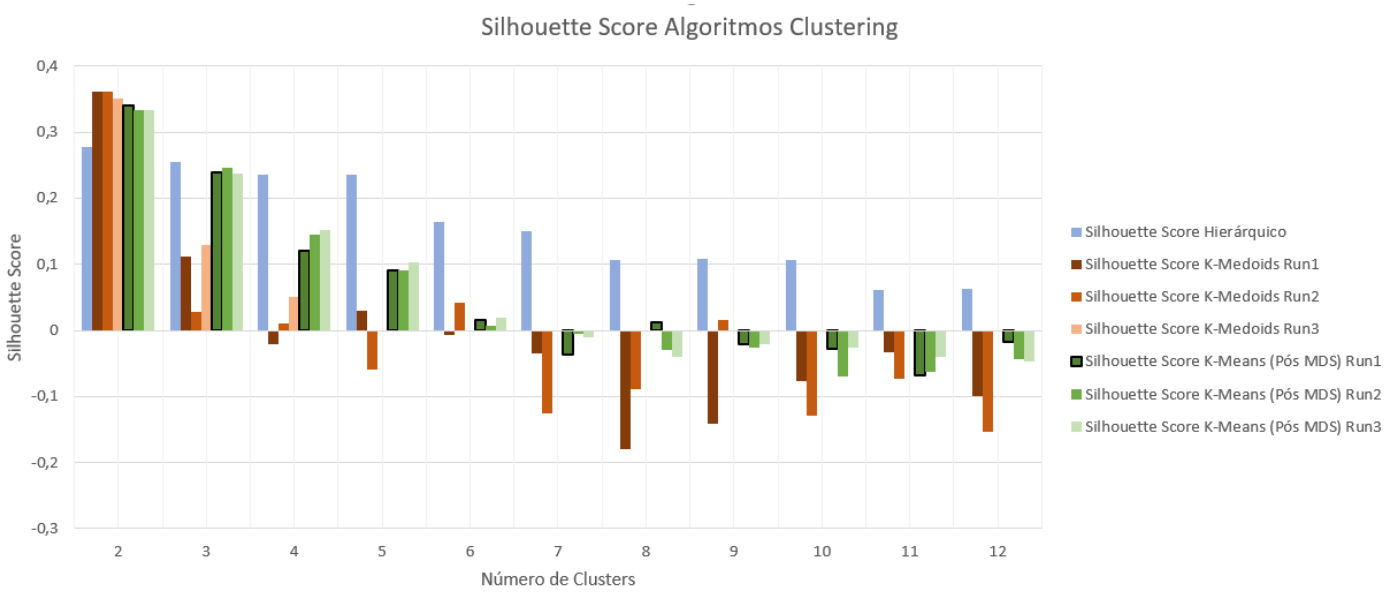


Figura 44 - *Silhouette Score* para os diferentes algoritmos

A primeira observação que se pode fazer relativamente ao número de *Clusters* é que todos tiveram a sua pontuação máxima para 2 *Clusters*, aproximadamente **0,35** no caso do *K-Medoids*. Sendo a range de score do método *Silhouette* entre -1 e 1 podemos aferir que ***'a nossa estrutura é fraca e pode ser artificial'*** (Anja Struyf, 2014) para 2 *Clusters*. Outro ponto importante a assinalar é o facto de o algoritmo Hierárquico ser o mais consistente a nível de resultado à medida que aumentam o número de *Clusters* sendo que os restantes caiem muito a partir de 3 *Clusters*. Relativamente ao *K-Medoids* é possível constatar que para mais três ou mais *Clusters*, há uma grande diferença de pontuação entre a *run1* e a *run2* o que mostra o quão sensível este algoritmo é aos *Medoids* iniciais fornecidos de forma *random*. O Algoritmo *K-Means* é sempre

muito consistente entre as *runs* ao longo do número de *Clusters* sendo a melhor pontuação também obtida para dois *Clusters*.

Sendo dois o número de *Clusters* com pontuação mais alta de forma maioritária para os três algoritmos, em particular a *run1* do algoritmo *K-Medoids*, será esta segmentação que utilizaremos para o resto da análise. Podemos utilizar o mapa obtido pelo *MDS* (**Figura 42**), para visualizar os dois *Clusters* no plano, gerados pelo *K-Medoids* e os respetivos *Medoids* finais indicados com uma seta vermelha (**Figura 45**).

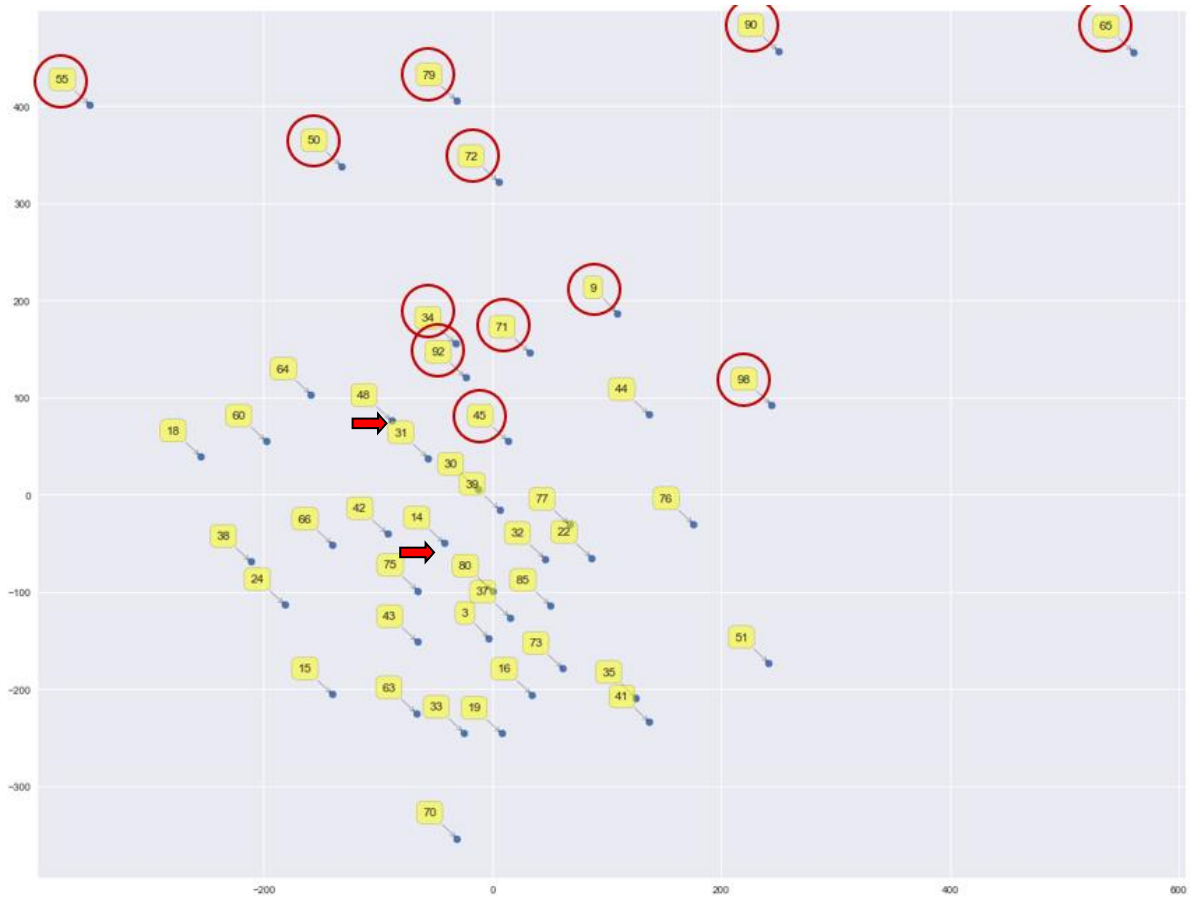


Figura 45 – Clusters obtidos pelo algoritmo *K-Medoids* em duas dimensões após aplicação do MDS

Fazendo uma comparação direta entre os *Clusters* formados pela *run1* do algoritmo *K-Medoids* e o algoritmo Hierárquico é possível observar que apenas 4 CPV's são assignados a *Clusters* diferentes o que significa 41 de 45 (91%), CPV's assignados aos mesmos *clusters*.

Os CPV's com *clusters* distintos são:

- **CPV 18 - Vestuário, calçado, malas e artigos de viagem, acessórios**
- **CPV 60 - Serviços de transporte (excl. transporte de resíduos)**
- **CPV 64 - Serviços postais e de telecomunicações**
- **CPV 98 - Outros serviços comunitários, sociais e pessoais**

Fazendo o mesmo exercício para comparar o algoritmo *K-Medoids* ao *K-Means* apenas 3 CPV's ficam assignados a *clusters* distintos sendo eles:

- **CPV 44** - Estruturas e materiais de construção; produtos auxiliares de construção (excetos aparelhos elétricos)
- **CPV 76** - Serviços relacionados com as indústrias do gás e do petróleo
- **CPV 45** - Construção

A **Tabela 11** mostra os resultados obtidos para os três algoritmos utilizados e os respetivos *Medoids* finais assinalados a amarelo associados à *run1* do *K-Medoids:*

- **CPV 39 - Mobiliário (incl. de escritório), acessórios, aparelhos domésticos (excl. iluminação) e produtos de limpeza**
- **CPV 92 - Serviços recreativos, culturais e desportivos**

		cluster_Hier (2 Clusters)	*cluster_Hier (3 Clusters)*	*K-Medoids (2 Clusters)*			*K-Medoids (3 Clusters)*			*K-Means (2 Clusters)*			*K-Means (3 Clusters)*		
ID	**CPV**	**Run1**	**Run1**	**Run1**	**Run2**	**Run3**	**Run1**	**Run2**	**Run3**	**Run1**	**Run2**	**Run3**	**Run1**	**Run2**	**Run3**
0	3	1	1	2	1	1	3	3	3	1	1	1	0	1	1
2	14	1	1	2	1	1	3	3	3	1	1	1	0	1	1
3	15	1	1	2	1	2	3	3	2	1	1	1	0	1	1
4	16	1	1	2	1	2	1	3	2	1	1	1	0	1	1
5	18	2	3	2	2	1	3	2	1	1	1	0	2	0	0
6	19	1	1	2	1	1	1	3	3	1	1	1	0	1	1
7	22	1	1	2	1	1	3	3	3	1	1	1	0	1	1
8	24	1	1	2	1	2	3	3	2	1	1	1	0	1	0
9	30	1	1	2	1	1	3	3	3	1	1	1	0	1	0
10	31	1	1	2	1	1	3	3	3	1	1	0	2	0	0
11	32	1	1	2	1	1	3	3	3	1	1	1	0	1	1
12	33	1	1	2	1	2	3	3	2	1	1	1	0	1	1
13	34	2	3	1	1	1	3	3	3	0	0	0	2	0	0
14	35	1	1	2	1	2	3	3	2	1	1	1	0	1	1
15	37	1	1	2	1	1	3	3	3	1	1	1	0	1	1
16	38	1	1	2	1	1	1	3	3	1	1	1	0	1	0
17	39	1	1	2	1	1	3	3	3	1	1	1	0	1	1
18	41	1	1	2	1	2	1	3	2	1	1	1	0	1	1
19	42	1	1	2	1	1	3	3	3	1	1	1	0	1	0
20	43	1	1	2	1	1	3	3	3	1	1	1	0	1	1
21	44	1	1	2	1	2	3	1	2	0	0	0	2	2	2
22	45	2	3	1	1	1	3	3	3	1	1	0	2	0	0
23	48	2	3	2	1	1	3	3	3	1	1	0	2	0	0
24	50	2	3	1	2	1	2	3	3	0	0	0	2	0	0
25	51	1	1	2	1	2	1	3	2	1	1	1	0	1	1
26	55	2	3	1	2	2	3	3	2	0	0	0	2	0	0
27	60	2	3	2	1	1	3	3	3	1	1	0	2	0	0
28	63	1	1	2	1	1	3	3	3	1	1	1	0	1	1
29	64	2	3	2	1	1	3	3	3	1	1	0	2	0	0
30	65	2	2	1	2	2	3	3	2	0	0	0	1	2	2
31	66	1	1	2	1	1	3	3	3	1	1	1	0	1	0
32	70	1	1	2	1	2	1	1	2	1	1	1	0	1	1
33	71	2	3	1	2	1	3	3	3	0	0	0	2	0	0
34	72	2	3	1	2	1	3	3	3	0	0	0	2	0	2
35	73	1	1	2	1	2	3	3	2	1	1	1	0	1	1
36	75	1	1	2	1	1	3	3	3	1	1	1	0	1	1
37	76	1	1	1	1	1	1	3	3	1	1	1	0	1	1
38	77	1	1	2	1	1	3	3	3	1	1	1	0	1	1
39	79	2	3	1	2	1	2	3	3	0	0	0	2	0	2
40	80	1	1	2	1	2	3	3	2	1	1	1	0	1	1
41	85	1	1	2	1	2	3	3	2	1	1	1	0	1	1
1	9	2	3	1	1	1	3	3	3	0	0	0	2	2	2
42	90	2	2	1	2	2	2	3	2	0	0	0	1	2	2
43	92	2	3	1	2	1	3	3	3	0	0	0	2	0	0
44	98	2	3	1	2	1	3	3	3	0	0	0	2	2	2

Tabela 11 - *Clusters* formados pelos três algoritmos

4.2.Análise dos clusters obtidos

Nesta seção é feita uma análise exploratória dos *Clusters* obtidos de modo a tentar caracteriza-los.

A **Tabela 12** mostra os *Clusters* finais gerados pelo algoritmo *K-Medoids*.

CPV	CPV Desc	Cluster
9	Produtos petrolíferos, combustíveis, electricidade e outras fontes de energia	C1
34	Equipamento e produtos auxiliares de transporte	C1
45	Construção	C1
50	Serviços de reparação e manutenção	C1
55	Serviços de hotelaria, restauração e comércio a retalho	C1
65	Serviços públicos	C1
71	Serviços de arquitectura, construção, engenharia e inspecção	C1
72	Serviços de TI: consultoria, desenvolvimento de software, Internet e apoio	C1
79	Serviços a empresas: direito, comercialização, consultoria, recrutamento, impressão e segurança	C1
90	Serviços relativos a águas residuais, resíduos, limpeza e ambiente	C1
92	Serviços recreativos, culturais e desportivos	C1
98	Outros serviços comunitários, sociais e pessoais	C1
3	Produtos da agricultura, da pesca, da silvicultura e afins	C2
14	Exploração mineira, metais de base e produtos afins	C2
15	Produtos alimentares, bebidas, tabaco e produtos afins	C2
16	Maquinaria agrícola	C2
18	Vestuário, calçado, malas e artigos de viagem, acessórios	C2
19	Materiais têxteis, de couro, de plástico e de borracha	C2
22	Material impresso e afins	C2
24	Produtos químicos	C2
30	Máquinas, equipamento e material de escritório e de informática, excepto mobiliário e pacotes de programas (software)	C2
31	Maquinaria, aparelhagem, equipamento e consumíveis eléctricos; iluminação	C2
32	Equipamento de rádio, televisão, comunicação, telecomunicações e afins	C2
33	Equipamento médico, medicamentos e produtos para cuidados pessoais	C2
35	Equipamento de segurança, combate a incêndios, polícia e defesa	C2
37	Instrumentos musicais, artigos de desporto, jogos, brinquedos, material para artesanato e actividades artísticas e acessórios	C2
38	Equipamento laboratorial, óptico e de precisão (exc. óculos)	C2
39	Mobiliário (incl. de escritório), acessórios, aparelhos domésticos (excl. iluminação) e produtos de limpeza	C2
42	Máquinas industriais	C2
43	Maquinaria para extracção mineira e pedreiras, equipamento de construção	C2
44	Estruturas e materiais de construção; produtos auxiliares de construção (excepto aparelhos eléctricos)	C2
48	Pacotes de software e sistemas de informação	C2
51	Serviços de instalação (excepto software)	C2
60	Serviços de transporte (excl. transporte de resíduos)	C2
63	Serviços anexos e auxiliares dos transportes; serviços de agências de viagens	C2
64	Serviços postais e de telecomunicações	C2
66	Serviços de finanças e seguros	C2
70	Serviços imobiliários	C2
73	Serviços de investigação e desenvolvimento e serviços de consultoria conexos	C2
75	Serviços relacionados com a administração pública, a defesa e a segurança social	C2
77	Serviços de agricultura, silvicultura, horticultura, aquicultura e apicultura	C2
80	Serviços de ensino e formação	C2
85	Serviços de saúde e acção social	C2

Tabela 12 - *Clusters* Finais *K-Medoids*

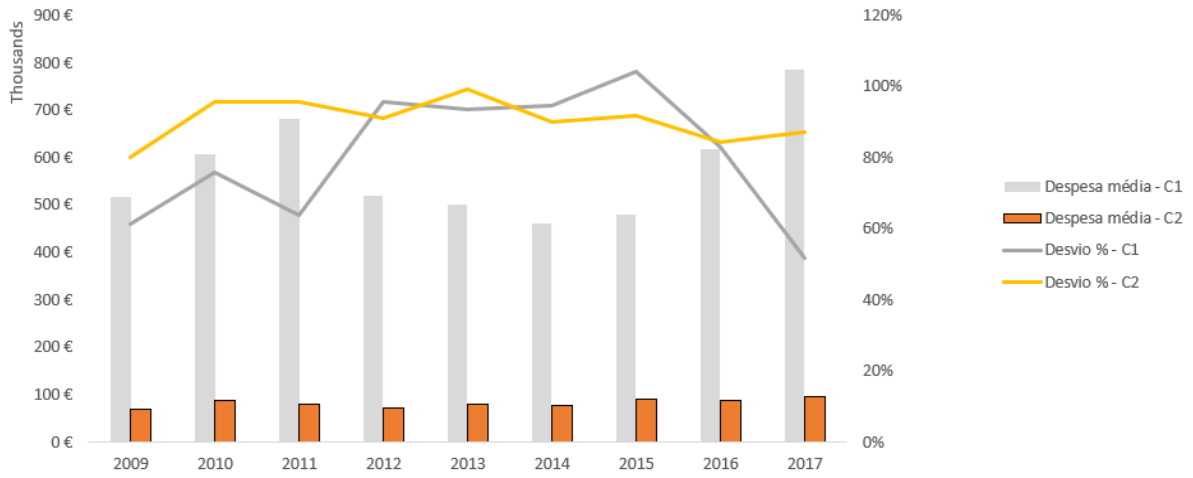


Figura 46 - Evolução média anual da despesa/desvio por *Cluster*

Como se pode verificar no gráfico ilustrado na **Figura 46**, a despesa média anual do *Cluster* 1 é claramente mais elevada que o *Cluster* 2 para todos os anos. Relativamente ao desvio percentual, o *Cluster* C1 apresenta um desvio global médio percentual de 76%, claramente inferior aos 90% do *Cluster* C2 o que evidencia que a segmentação criada pelo algoritmo *K-Medoids* criou uma separação entre tipos de contrato visível na métrica do desvio médio. Ainda podemos constatar que não há uma associação direta entre a despesa média e desvio.

5.Conclusão

5.1.Principais Resultados

Este trabalho teve dois grandes objetivos, a extração automática dos contratos públicos do site http://www.base.gov.pt/ através de um *crawler* desenvolvido em *Python* e a posterior segmentação dos mesmos em termos de desvios monetários face à *trend line* natural da relação entre dimensão populacional e despesa municipal.

O desenvolvimento do *crawler* neste projeto foi um fator determinante para a posterior análise, podendo ser adaptado a outros cenários. A extração de dados *html* da *web* e o seu posterior tratamento permite a integração de nova informação em qualquer objeto de estudo. As funções fornecidas pelo *Python* relativas a *web Scraping* relevaram-se fundamentais para a extração dos contratos nomeadamente a biblioteca ***Beautiful Soup*.** O facto de terem sido extraídos os contratos em bruto do gov.pt requereu posteriormente um trabalho de transformação e tratamento de dados de forma a criar um *dataset* final para análise. Destaca-se a utilização das bibliotecas ***Pandas* e *NumPy*** que têm inúmeras funções e métodos de manipulação de tabelas e *Arrays*. Sendo um objetivo deste projeto a utilização de análise estatística e técnicas de *DataMining*, o *Python* revelou ser uma linguagem de programação muito completa com diversos algoritmos disponíveis para análise quer estatística, quer de segmentação, com destaque para os módulos ***Statsmodels* e *Scikit-learn***.

Relativamente à análise, é possível concluir que há claramente dois grupos de CPV's com desvios monetários face à tendência natural. Foi possível retirar esta conclusão analisando os resultados obtidos pelo método de avaliação de *Clusters*, *Silhouette Score,* que para as três técnicas de *Clustering* usadas teve uma pontuação média de **0,34**, claramente superior às pontuações obtidas para as partições com mais de dois *Clusters*. O algoritmo que melhor conseguiu separar os CPV's foi o *K-Medoids* que obteve um *Silhouete Score* de **0,36** para uma das ***runs.*** Outro ponto a assinalar é o facto deste valor ser consistente já que foram executadas várias runs com *Medoids* iniciais gerados de forma aleatória e o valor manteve-se aproximado.

Embora tenhamos optado pelo uso do algoritmo *K-Medoids* foi possível visualizar que a estratégia de redução de dimensionalidade para duas dimensões gerada pelo algoritmo *MDS* seguida da *Clusterização* usando o algoritmo *K-Means*, conseguiu separar de forma muito equivalente ao *K-Medoids* os CPV's, classificando 39 dos 45 CPV's nos mesmos *Clusters*, ou seja 87%. Esta

consistência entre ambos os métodos, mostra o quão importante é o algoritmo *MDS* no que toca à redução de dimensionalidade e consequente possibilidade de permitir visualizar como estão distribuídos os CPV's no plano.

Relativamente aos dois *Clusters* formados durante a análise podemos observar que um deles tem um consumo médio global bastante superior embora o desvio médio seja inferior o que não permite concluir a possibilidade de haver uma relação direta entre gasto e desvios médios por município/ano. Ainda é possível concluir que ambos os *clusters* têm um desvio médio positivo. Este estudo permitiu concluir que a identificação de clusters pode simplificar a análise futura dos padrões de despesa bem como o entendimento de uma classificação, que apesar de todo o detalhe é complexa.

5.2.Sugestões e Limitações

A análise deste projeto teve como ponto principal a obtenção de Clusters de CPV's, o que foi conseguido para dois grupos, com um grau de fiabilidade de 0,36 através do método de avaliação de Clusters, Silhouette Score, o que segundo Anja Struyf e Mia Hubert 'a nossa estrutura é fraca e pode ser artificial'. O facto da segmentação obtida ter tido como ponto de partida os desvios da relação entre despesa municipal e dimensão populacional através de uma relação log-log e algumas destas não terem sido explicativas para determinados CPV's-Ano, pode ter sido um fator determinante desta eventual 'fraca estrutura'. Um fator possivelmente explicativo, poderá ser o facto do número de contratos lançados ser sempre superior ano após ano, sugerindo que, muitas despesas não foram lançadas em anos iniciais, causando uma elevada percentagem de missing data ou municípios/anos/CPV's sem imputação.

A criação de dois grupos de CPV's poderá ser importante para detetar futuros desvios anómalos de despesa ao nível município ano, partindo do princípio que as condições se mantêm equivalentes. Outra utilização prática é a capacidade de poderem ser classificados contratos a um nível mais baixo, isto é, com mais de dois algarismos. Isto é possível devido à segmentação ter sido feita com o algoritmo K-Medoids que gerou dois Medoids finais representativos dos respetivos Clusters, neste caso em particular, CPV's concretos.

Como sugestão para futuros estudos, a relação log-log mencionada anteriormente poderá ser melhorada com a introdução de mais variáveis que influenciam o gasto municipal para além da dimensão populacional, como por exemplo, idade, rendimento ou dimensão/localização geográfica. Ainda poderá ser feita uma análise análoga à efetuada neste estudo, determinando clusters de Municípios em vez de clusters de tipos de despesa.

6.Bibliografia

Afonso, A., & Fernandes, S. (2003). Efficiency of Local Government Spending: Evidence for the Lisbon Region.

Almeida, V. (2001). O Estado, A Economia e as Despesas Públicas em Portugal. (APAPP, Ed.)

Anja Struyf, M. H. (1997). Clustering in an Object-Oriented Environment. *Journal of Statistical Software; 1997; Vol. 1*.

Bradford, D., Malt, R., & Oates, W. (1969). The Rising Cost of Local Public. *National Tax Journal*, 185-202.

Chatterjee, S. H. (2000). Chatterjee, S, Hadi, A.S., Price, B. Regression Analysis by Example , 3rd edition, Wiley, New York, p104, (2000). *Regression Analysis by Example , 3rd edition, Wiley, New York, p104, (2000).*

Cousineau, D. &. (2010). Outlier detection and treatment: a review. *International Journal of Psychological Research, ISSN 2011-7922, Vol. 3, Nº. 1, 2010, pags. 58-67. 3. .*

Dhakal, C. (2017). Dealing With Outliers And Influential Points While Fitting Regression. *Journal of Institute of Science and Technology. 22. 61. 10.3126/jist.v22i1.17741. .*

Eirola E, D. G. (2013). Distance estimation in numerical data sets with missing values. Inform Sci 2013; 240: 115–128. *Inform Sci 2013; 240: 115–128.*

Farrel, M. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society*.

Ferrara, E., Fiumara, P., & Baumgartner, R. (July de 2012). Web Data Extraction, Applications and Techniques: A Survey. *Knowledge-Based Systems*.

Fisher, R. (1996). *State and local Public Finance.* Irwin.

Friedman, M., & Friedman, R. D. (1962). *Capitalism and Freedom.* Chicago: University of Chicago Press.

Ian H. Witten, F. E. (s.d.). Data mining : practical machine learning tools and techniques.—3rd ed. .

Keen, B. A. (2019). *http://benalexkeen.com/feature-scaling-with-scikit-learn/*.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money.* London: Macmillan.

Klein, N. (2014). *This Changes Everything: Capitalism vs the Climate. Simon & Schuster.*

Kuljanin, A., & Klipstein, C. (2017). *Consultancy Services for Common Procurement Vocabulary expert group.*

Lovell, C. (1993). *Production Frontiers and Productive Efficiency.* Oxford University Press.

Lovell, C. (2000). Measuring Efficiency in the Public Sector. Em J. Blank, *Public Provision and Performance.*

McQueen, J. (1967). Some methods for classification and analysis of multivariate observation. . *L. Le Cam and J.Neyman (Eds.), 5th Berkeley Symp. Math. Stat. Prob., 1,pp281-297.* .

Onukwugha, C. (2018). Data Mining Application in Credit Card Fraud Detection System.

Padhy N., M. P. (2012). The Survey of Data Mining Applications and Feature Scope; International Journal of Computer Science, Engineering and Information Technology .

Paea, S. a. (2018). Information Architecture (IA): Using Multidimensional Scaling (MDS) and K-Means Clustering Algorithm for Analysis of Card Sorting Data. *Paea, Sione and Baird, Ross (2018) Information Architecture (IA): Using multidimensional scaling (MDS) and k-means cluJournal of Usability Studies, 13 (3). pp. 138-157. ISSN 1931-3357.*

pordata. (s.d.). *Administrações Públicas: despesas, receitas e défice/excedente público, em % do PIB.* Obtido de https://www.pordata.pt/: https://www.pordata.pt/Europa/Administra%C3%A7%C3%B5es+P%C3%BAblicas+despesas++receitas+e+d%C3%A9fice+excedente+p%C3%BAblico++em+percentagem+do+PIB-1762

Raykov, Y. P., Boukouvalas, A., & Baig, F. a. (2016). What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. *Byung-Jun Yoon. PLOS ONE 11, no. 9 (September 2016): e0162259. © 2016 Raykov et al*.

Rui Xu. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks Vol. 16, No. 3, May 2005.*

Santhana Chaimontree, K. A. (2010). Best Clustering Configuration Metrics: Towards. *ADMA (1) 2010: 48-59.*

Santos, A. (s.d.). A Evolução das Despesa Públicas em Portugal – aspectos de Longo Prazo. Em 1984, *Estudos de Economia* (pp. 487 - 501).

Saukar, A. V., Kedar, P. G., & Gode, S. A. (April de 2018). An Overview On Web Scraping Techniques And Tools. *International Journal on Future Revolution in Computer Science & Communication Engineering, 4*(4), 363 – 367.

Tanzi, V., & Schuknecht, L. (2000). *Public Spending in the 20th Century- A Global Perspective.* Cambridge: Cambridge University Press.

Trevor Hastie, R. T. (2001). *The elements of statistical learning: data mining, inference, and prediction.* Springer.

Usman, I. B. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology 6(17): 3299-3303, 2013.*

Vargiu, E., & Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research, 2*.

Wooldridge, J. M. (2013). Introductory Econometrics: A Modern Approach, fifth edition. Cincinnati, OH: South-Western.

Xi Hang Cao, I. S. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics. 2016; 17(1): 359.*

Yamaganti, R., & Sikharam, U. M. (Abril de 2015). Data Warehousing Concepts Using ETL Process for Social Media Data Extraction. *International Journal of Scientific & Engineering Research, 6*(4).

Yang, X.-S. (2019). *Introduction to Algorithms for Data Mining and Machine Learning.* Academic Press.

7.Anexos

Anexo 1 - Identificação e descrição dos contratos relativos aos 2 primeiros algarismos (Divisões)[7]

CPV	CPV descrição
03	Produtos da agricultura, da pesca, da silvicultura e afins
09	Produtos petrolíferos, combustíveis, electricidade e outras fontes de energia
14	Exploração mineira, metais de base e produtos afins
15	Produtos alimentares, bebidas, tabaco e produtos afins
16	Maquinaria agrícola
18	Vestuário, calçado, malas e artigos de viagem, acessórios
19	Materiais têxteis, de couro, de plástico e de borracha
22	Material impresso e afins
24	Produtos químicos
30	Máquinas, equipamento e material de escritório e de informática, excepto mobiliário e pacotes de programas (software)
31	Maquinaria, aparelhagem, equipamento e consumíveis eléctricos; iluminação
32	Equipamento de rádio, televisão, comunicação, telecomunicações e afins
33	Equipamento médico, medicamentos e produtos para cuidados pessoais
34	Equipamento e produtos auxiliares de transporte
35	Equipamento de segurança, combate a incêndios, polícia e defesa
37	Instrumentos musicais, artigos de desporto, jogos, brinquedos, material para artesanato e actividades artísticas e acessórios
38	Equipamento laboratorial, óptico e de precisão (exc. óculos)
39	Mobiliário (incl. de escritório), acessórios, aparelhos domésticos (excl. iluminação) e produtos de limpeza
41	Água captada e tratada.
42	Máquinas industriais
43	Maquinaria para extracção mineira e pedreiras, equipamento de construção
44	Estruturas e materiais de construção; produtos auxiliares de construção (excepto aparelhos eléctricos)
45	Construção
48	Pacotes de software e sistemas de informação
50	Serviços de reparação e manutenção
51	Serviços de instalação (excepto software)
55	Serviços de hotelaria, restauração e comércio a retalho
60	Serviços de transporte (excl. transporte de resíduos)
63	Serviços anexos e auxiliares dos transportes; serviços de agências de viagens
64	Serviços postais e de telecomunicações
65	Serviços públicos
66	Serviços de finanças e seguros
70	Serviços imobiliários
71	Serviços de arquitectura, construção, engenharia e inspecção
72	Serviços de TI: consultoria, desenvolvimento de software, Internet e apoio
73	Serviços de investigação e desenvolvimento e serviços de consultoria conexos
75	Serviços relacionados com a administração pública, a defesa e a segurança social
76	Serviços relacionados com as indústrias do gás e do petróleo
77	Serviços de agricultura, silvicultura, horticultura, aquicultura e apicultura
79	Serviços a empresas: direito, comercialização, consultoria, recrutamento, impressão e segurança
80	Serviços de ensino e formação
85	Serviços de saúde e acção social
90	Serviços relativos a águas residuais, resíduos, limpeza e ambiente
92	Serviços recreativos, culturais e desportivos
98	Outros serviços comunitários, sociais e pessoais

[7] https://www.espap.pt/Documents/servicos/compras/CPV_2008.xls

Anexo 2 – Excerto do anuário financeiro dos Municípios Portugueses 2017

Tipo	#	Designação	Particip. Mun.	Património líquido	Resultados Líquidos	Dívidas a terceiros		Índice. Dívida Total	Dívidas de terceiros	N.º de trab.	N.º de hab.
						Empréstimos	Outras				
CM	G	Almada	–	332.174.960	1.283.294	27.775.105	5.549.500	41,1%	6.978.791	1.589	169.152
SMAS		Almada	100%	47.620.781	-3.189.808	0	9.621.846	–	7.175.810	478	–
EM SA		ECALMA-Estacionamento e Circulação	100%	291.447	1.157	0	392.292	–	51.731	–	–
Grupo		Contas consolidadas	–	379.510.964	-2.004.249	27.775.105	6.346.747	–	6.841.329	–	169.152
CM	P	Almeida	–	30.858.081	229.860	2.178.100	1.198.985	30,4%	605.658	153	6.062
CM	M	Almeirim	–	50.759.374	-1.773.696	5.161.666	602.188	44,5%	224.720	235	22.782
CM	P	Almodôvar	–	41.771.570	38.576	3.881.050	944.354	47,3%	167.927	202	6.813
Grupo		Contas consolidadas	–	41.771.570	32.133	3.881.050	944.354	–	167.927	–	6.813
CM	P	Alpiarça	–	25.887.723	-1.329.661	6.544.676	2.012.283	155,1%	469.160	151	7.155
Grupo		Contas consolidadas	–	23.875.330	-990.341	6.676.941	1.794.015	–	298.344	–	7.155
CM	P	Alter do Chão	–	26.834.327	-774.912	882.491	904.890	31,6%	895.523	135	3.229
CM	P	Alvaiázere	–	36.228.956	-1.246.577	3.267.206	530.991	56,0%	157.934	86	6.710
CM	P	Alvito	–	18.461.494	598.574	716.592	262.604	24,2%	730.869	110	2.459
CM	G	Amadora	–	319.586.343	12.592.442	21.603.622	6.095.430	33,6%	5.550.440	1.693	179.942
SIMAS		Oeiras e Amadora	50,0%	160.560.310	8.040.679	0	10.204.826	–	10.433.876	394	–
EM		Amadora Inovation, E. M. Unipessoal, Lda.	100%	389.453	-327.042	0	230.982	–	419.118	–	–
Grupo		Contas consolidadas	–	400.206.960	15.087.388	21.603.622	11.327.952	–	10.687.259	–	179.942
CM	M	Amarante	–	91.845.002	-1.363.311	10.762.231	3.608.018	51,7%	1.050.154	559	53.614
CM	P	Amares	–	26.459.860	-126.993	5.230.667	2.246.843	63,6%	408.790	198	18.147
CM	M	Anadia	–	87.281.079	-977.149	4.662.324	696.136	30,8%	615.615	232	27.576
Grupo		Contas consolidadas	–	87.278.668	-1.151.499	4.762.324	745.978	–	769.692	–	27.576
CM	M	Angra do Heroísmo	–	91.900.555	1.477.314	13.469.902	1.886.581	108,9%	8.373.295	220	34.105
SMAS		Angra do Heroismo	100%	7.518.757	153.770	0	1.737.753	–	572.878	143	–
EM		TERAMB	60,0%	28.972.052	-517.501	2.171.906	5.990.061	–	801.983	–	–
Grupo		Contas consolidadas	–	Si	Si	Si	Si	-	Si	-	34.105
CM	P	Ansião	–	41.319.066	484.450	5.687.358	923.375	81,1%	184.094	102	12.270
Grupo		Contas consolidadas	–	41.319.810	484.526	4.434.926	2.175.807	–	184.094	–	12.270
CM	M	Arcos de Valdevez	–	77.316.426	22.907	4.191.286	3.578.515	34,5%	1.288.491	296	21.144
Grupo		Contas consolidadas	–	77.896.689	35.300	4.191.286	3.578.515	–	1.288.491	–	21.144
CM	P	Arganil	–	37.302.503	-153.970	2.250.000	1.557.296	32,6%	395.606	186	11.181
CM	P	Armamar	–	21.485.149	-93.297	3.898.611	1.892.225	78,9%	74.573	172	5.838
CM	M	Arouca	–	44.794.534	1.234.259	2.072.429	1.232.795	21,7%	3.873	175	21.039
CM	P	Arraiolos	–	38.393.874	51.136	3.471.691	710.191	60,2%	204.784	139	6.999
CM	P	Arronches	–	23.120.023	529.507	868.276	211.818	22,0%	32.671	98	2.910
CM	P	Arruda dos Vinhos	–	22.134.199	266.260	3.622.474	1.943.679	56,2%	541.342	203	14.925
CM	M	Aveiro	–	120.281.482	7.136.189	76.227.213	29.599.487	229,6%	8.381.829	585	77.630
EM		Aveiro Expo - Parque de Exposições	51,0%	26.208	193.797	194.896	493.451	–	608.693	–	–
EM		EMA - Estádio Municipal de Aveiro	100%	26.383.743	-241.854	195.704	4.904.067	–	868.002	–	–

Anexo 3 - Associação Município – NIF

Município	nif_mun
Almada	500051054
Sintra	500051062
Lisboa	500051070
Alcanena	500745773
Oeiras	500745943
Cuba	500832935
Caminha	500843139
Tavira	501067191
Marco de Canaveses	501073655
Penafiel	501073663
Barrancos	501081216
Águeda	501090436
Felgueiras	501091823
Amarante	501102752
Serpa	501112049
Entroncamento	501120149
Idanha-a-Nova	501121030
Lousã	501121528
Torre de Moncorvo	501121536
Aljustrel	501122486
Oliveira do Bairro	501128840
Mora	501129103
Guarda	501131140
Alter do Chão	501132872
Alpiarça	501133097
Castro Verde	501135960
Valongo	501138960
Castelo Branco	501143530
Portalegre	501143718
Trancoso	501143726
Vidigueira	501143734
Sesimbra	501144218
Arronches	501155996
Vinhais	501156003
Santa Maria da Feira	501157280
Espinho	501158740
Fronteira	501162941
Marvão	501170162
Campo Maior	501175229
Vendas Novas	501177256
Sardoal	501181857
Estarreja	501190082

Chaves	501205551
Mourão	501206639
Ferreira do Zêzere	501216839
Caldas da Rainha	501222634
Ferreira do Alentejo	501227490
Arraiolos	501258027
Mangualde	501262997
Elvas	501272968
Montemor-o-Velho	501272976
Almeirim	501273433
Condeixa-a-Nova	501275380
Ourém	501280740
Alvito	501288120
Batalha	501290206
Setúbal	501294104
Anadia	501294163
Loures	501294996
Chamusca	501305564
Figueira da Foz	501305580
Alenquer	501305734
Matosinhos	501305912
Porto	501306099
Oliveira de Frades	501306234
Ovar	501306269
Santo Tirso	501306870
Redondo	501834117
Monção	501937471
Loulé	502098139
Santiago do Cacém	502130040
Alcácer do Sal	502150319
Paços de Ferreira	502173297
Torres Vedras	502173653
Moura	502174153
Mafra	502177080
Lourinhã	502177101
Sines	502563010
Abrantes	502661038
Castelo de Paiva	502678917
Sever do Vouga	502704977
Avis	502789824
Montijo	502834846
São Brás de Alportel	503219924
Mértola	503279765
Albufeira	503539473
Borba	503956546

Odivelas	504293125
Trofa	504296434
Évora	504828576
Beja	504884620
Mêda	505161974
Lagos	505170876
Leiria	505181266
Cascais	505187531
Arcos de Valdevez	505211696
Lousada	505279460
Arruda dos Vinhos	505307685
Portimão	505309939
Odemira	505311313
Cabeceiras de Basto	505330334
Covilhã	505330768
Vila Nova de Gaia	505335018
Vila Nova de Poiares	505371600
Proença-a-Nova	505377802
Maia	505387131
Sobral de Monte Agraço	505410850
Amadora	505456010
Barcelos	505584760
Porto de Mós	505586401
Melgaço	505592940
Fornos de Algodres	505592959
Rio Maior	505656000
Ponte da Barca	505676770
Cadaval	505763621
Marinha Grande	505776758
Vila do Conde	505804786
Aveiro	505931192
Aljezur	505932512
Santarém	505941350
Guimarães	505948605
Vizela	505985217
Figueira de Castelo Rodrigo	505987449
Viana do Castelo	506037258
Cantanhede	506087000
Montalegre	506149811
Viana do Alentejo	506151174
Seixal	506173968
Palmela	506187543
Penamacor	506192164
Portel	506196445
Bragança	506215547

Fundão	506215695
Oliveira de Azeméis	506302970
Olhão	506321894
Pombal	506334562
Resende	506349381
Vila Real	506359670
Coimbra	506415082
Gouveia	506510476
São João da Madeira	506538575
Figueiró dos Vinhos	506546381
Estremoz	506556590
Golegã	506563774
Lamego	506572218
Faro	506579425
Tabuaço	506601455
Ansião	506605930
Alvaiázere	506605949
Torres Novas	506608972
Montemor-o-Novo	506609553
Nisa	506612287
Góis	506613399
Vila Viçosa	506613461
Vila Franca de Xira	506614913
Esposende	506617599
Miranda do Corvo	506624200
Almeida	506625419
Vimioso	506627888
Póvoa de Lanhoso	506632920
Paredes de Coura	506632938
Manteigas	506632946
Santa Comba Dão	506637441
Vila Verde	506641376
Vila Velha de Ródão	506642798
Alfândega da Fé	506647498
Penedono	506651541
Paredes	506656128
Penacova	506657957
Vieira do Minho	506659682
Crato	506659968
Vila Nova de Famalicão	506663264
Moimenta da Beira	506664686
Carrazeda de Ansiães	506666018
Barreiro	506673626
Benavente	506676056
Seia	506676170

Carregal do Sal	506684920
Cinfães	506693651
Belmonte	506695956
Vila Flor	506696464
Viseu	506697320
Macedo de Cavaleiros	506697339
Castro Daire	506716210
Coruche	506722422
Mira	506724530
Valença	506728897
Vila do Bispo	506730573
Castanheira de Pêra	506731324
Vale de Cambra	506735524
Tomar	506738914
Póvoa de Varzim	506741400
Tarouca	506753905
Salvaterra de Magos	506755150
Vouzela	506770664
Alcoutim	506772446
Alandroal	506772527
Penela	506778037
Cartaxo	506780902
Albergaria-a-Velha	506783146
São Pedro do Sul	506785815
Pinhel	506787249
Alcochete	506788490
Moita	506791220
Murtosa	506791238
Mealhada	506792382
Penalva do Castelo	506792404
Castelo de Vide	506796035
Amares	506797627
Bombarral	506800580
Castro Marim	506801969
Óbidos	506802698
Lagoa	506804240
Ponte de Sor	506806456
Miranda do Douro	506806898
Tábua	506806944
Arouca	506808122
Aguiar da Beira	506809307
Vila Nova de Paiva	506809323
Sousel	506809560
Vila Pouca de Aguiar	506810267
Sabugal	506811662

Pampilhosa da Serra	506811883
Ponte de Lima	506811913
Peniche	506812820
Mação	506814343
Almodôvar	506816184
Ribeira de Pena	506818098
Oliveira do Hospital	506818829
Silves	506818837
Azambuja	506821480
Tondela	506822680
Grândola	506823318
Oleiros	506824152
Sabrosa	506824942
Constância	506826546
Monchique	506826961
Santa Marta de Penaguião	506829138
Vila Nova de Foz Côa	506829197
Peso da Régua	506829260
Vila Real de Santo António	506833224
Arganil	506833232
Nelas	506834166
Mesão Frio	506840328
Fafe	506841561
Armamar	506843190
Gondomar	506848957
Celorico da Beira	506849635
Mogadouro	506851168
Sernancelhe	506852032
Baião	506854299
Mortágua	506855368
Alijó	506859487
Murça	506862763
Gavião	506865517
Monforte	506873412
Alcobaça	506874249
Valpaços	506874320
Ourique	506876330
Mirandela	506881784
Sátão	506882713
Celorico de Basto	506884929
Freixo de Espada à Cinta	506884937
Boticas	506886964
São João da Pesqueira	506892646
Vila Nova de Cerveira	506896625
Vila Nova da Barquinha	506899250

Braga	506901173
Terras de Bouro	506907619
Vagos	506912833
Ílhavo	506920887
Vila de Rei	506932273
Sertã	506963837
Mondim de Basto	506967107
Pedrógão Grande	507011937
Nazaré	507012100
Reguengos de Monsaraz	507040589
Soure	507103742
Funchal	511217315
Câmara de Lobos	511233620
Calheta [R.A.M.]	511233639
Ponta do Sol	511235461
Ribeira Brava	511236417
Porto Santo	511236425
Porto Moniz	511239068
Machico	511239440
Santana	511239980
São Vicente	511240112
Santa Cruz	511244681
Ponta Delgada	512012814
Ribeira Grande	512013241
Nordeste	512042659
Vila Franca do Campo	512043701
Vila da Praia da Vitória	512044023
Angra do Heroísmo	512044040
Vila do Porto	512063770
Povoação	512065047
Corvo	512065837
Santa Cruz da Graciosa	512069760
Madalena	512070946
Horta	512073821
Calheta [R.A.A.]	512074089
Lajes do Pico	512074143
Lagoa [R.A.A.]	512074410
São Roque do Pico	512074771
Lajes das Flores	512074836
Velas	512075506
Santa Cruz das Flores	512079110

Anexo 4 - Associação Município - EM

Município	nif_mun	Empresa	nif_em	Tipo
Abrantes	502661038	Ambientabrantes	680017542	SMA
Aguiar da Beira	506809307	ABTT - Aguiar da Beira Termas e	508310709	EEM
Albergaria-a-Velha	506783146	Albergaria-a-Velha	680037888	SMAS
Alcobaça	506874249	Alcobaça	680014942	SMAS
Alfândega da Fé	506647498	Alfandegado - Desenvolvi- mento	502649631	EM
Alfândega da Fé	506647498	EDEAF - EM Desenvolvimento	506666573	EM
Almada	500051054	ECALMA - Estacionamen- to e	507001206	EM
Almada	500051054	SMAS de Almada	680017763	SMAS
Almeida	506625419	Almeida Municipia	507378890	EEM
Almeirim	501273433	ECOLEZIRIA - Tratamento de	504871650	EIM
Amadora	505456010	Amadora Inovation, E. M. Unipessoal,	504746383	EM
Amadora	505456010	Escola Intercultural das Profi ssões e do	504746383	EM
Amadora	505456010	Amadora e Oeiras	680015019	SIMAS
Anadia	501294163	WRC - Agência de Desenvolvimento	506053628	SA
Anadia	501294163	Anadia	600071871	SMAS
Angra do Heroísmo	512044040	Culturangra	512099499	EEM
Angra do Heroísmo	512044040	TERAMB	509620515	EM
Angra do Heroísmo	512044040	Angra do Heroismo	680018140	SMAS
Aveiro	505931192	Aveiro Expo - Parque de Exposições	507095677	EM
Aveiro	505931192	Moveaveiro	507190327	EM
Aveiro	505931192	Teatro Aveirense	507327985	EM
Aveiro	505931192	Aveiro	680012842	SMAS
Azambuja	506821480	EMIA - Infraestrutu- ras de Azambuja	506980049	EM
Barcelos	505584760	EMDB - Empresa Municipal de	504623842	EM
Barcelos	505584760	EMEC - EM de Educação e Cultura	504635417	EM
Barcelos	505584760	Empresa Municipal de Desportos	504623842	EM
Barcelos	505584760	Empresa Municipal de Educação e	504635417	EM
Barreiro	506673626	Barreiro	680015574	SMTC
Barreiro	506673626	Transp. Colectivos do Barreiro	680015574	SMTC
Batalha	501290206	IserBatalha - Gestão de Equip Urb,	504825461	EM
Beja	504884620	EMAS - Água e Sanea- mento	505991527	EM
Beja	504884620	Inovobeja - EM de Desenvolvimento	508999650	EM
Braga	506901173	AGERE - Águas, Efl uentes e Resíduos	504807692	EM
Braga	506901173	BRAGAHABIT - EM de Habitação	504537784	EM
Braga	506901173	InvestBraga - Agência p/ dinamiz.	504807706	EM
Braga	506901173	TUB - Transportes Ur- banos	504807684	EM
Braga	506901173	Teatro Circo de Braga	500463964	EM SA
Braga	506901173	BRAVAL - Valorização e Tratamento	503730947	SA
Cabeceiras de	505330334	Basto Solidário	508118174	EM
Cabeceiras de	505330334	Emunibasto	506417883	EM
Caldas da Rainha	501222634	Caldas da Rainha	680001069	SMAS
Calheta (R. A. M.)	511233639	Empreendimenntos Sol- Calheta	511272049	EM

Campo Maior	501175229	CampoMaior XXI	507745396	EM
Cantanhede	506087000	INOVA - Desenvolvimento Económico	506091481	EM
Cartaxo	506780902	Rumo 2020	507892283	EM
Cascais	505187531	Arcascais - Emp. Gestora do Aeródromo	507328230	EM
Cascais	505187531	EMAC - Ambiente	507396081	EM
Cascais	505187531	EMGHA - Emp. de Gestão do Parque	504538314	EM
Cascais	505187531	ESUC - Emp. De Serviços Urbanos de	504853635	EM
Cascais	505187531	Fortaleza de Cascais	507456300	EM
Cascais	505187531	Cascais Dinâmica	503589780	EM SA
Cascais	505187531	CASCAIS ENVOLVENTE – Gestão	504538314	EM SA
Cascais	505187531	Cascais Próxima	504853635	EM SA
Cascais	505187531	ETE - Empresa de Turismo do Estoril	503589780	EM SA
Castanheira de Pera	505187531	Prazilandia -Turismo e Ambiente	506579794	EM
Castanheira de Pera	506731324	Prazilândia -Turismo e Ambiente	506579794	EM
Castanheira de Pêra	506731324	Prazilandia - Turismo e Ambiente	506579794	EM
Castanheira de Pêra	506731324	RIBEIRAPERA	501452303	EM SA
Castelo Branco	501143530	NATURTEJO Empresa de Turismo	506836860	EIM
Castelo Branco	501143530	ALBIGEC - Empresa de Gestão de	505715449	EM
Castelo Branco	501143530	Terras da Beira Baixa	509614531	EM SA
Castelo Branco	501143530	Castelo Branco	680031065	SMAS
Castro Marim	506801969	NOVBAESURIS – EM Gestão e	508926645	EM SA
Celorico de Basto	506884929	Qualidade de Basto - Empresa para o	504695436	EM
Celorico de Basto	506884929	Qualidade de Basto	504695436	EM
Chaves	501205551	GEMC - Gestão de Equipam. município	506695018	EM SA
Cinfães	506693651	Quinta de Tuberais - Ensino Profi	504615858	EM
Coimbra	506415082	AC AGUAS DE COIMBRA	506566307	EEM
Coimbra	506415082	TC - Turismo de Coimbra	507135407	EM
Coimbra	506415082	Coimbra Inovação Parque	506787729	EM SA
Coimbra	506415082	WRC - Agência de Desenvolvimento	506053628	SA
Coimbra	506415082	Transportes Urbanos de Coimbra	680015965	SMTC
Coimbra	506415082	PRODESO - Ensino Profi s- sional, Lda.	502675870	Soc.
Covilhã	505330768	ICOVI – Infraestrut.	508282322	EEM
Covilhã	505330768	A D C - Águas da Covilhã	507611977	EM
Covilhã	505330768	Nova Covilhã, SRU - So- ciedade de	507291832	EM
Covilhã	505330768	Parkurbis - Parque de Ciência e	505456176	SA
Cuba	500832935	Centro de Estudos Diogo Dias Melgaz	508581303	Soc.
Esposende	506617599	EAMB - Esposende Ambiente	507068076	EEM
Esposende	506617599	ESPOSENDE 2000	503879614	EEM
Évora	504828576	HABEVORA - Gestão Habitacional	507013212	EEM
Évora	504828576	Évora Viva, SRU - Sociedade de	500697884	EM
Évora	504828576	Mercado Municipal de Évora	507013212	EM
Évora	504828576	SITEE - Sistema Integrado de Trans-	504878620	EM
Faro	506579425	FAGAR - Faro Ges- tão de Aguas e	507142217	EM
Faro	506579425	AMBIFARO - Agência para o	503714593	SA
Faro	506579425	Mercado Municipal de Faro	504497782	SA

Faro	506579425	Teatro Municipal de Faro	506971635	SM
Felgueiras	501091823	ACLEM - Arte Cultura e Lazer Empresa	507974530	EM
Felgueiras	501091823	EPF - Ensino Profi ssional de Felgueiras	504575848	Soc.
Figueira da Foz	501305580	Figueira Domus - Gestão de habitação	501460888	EM
Figueira da Foz	501305580	Figueira Paraindustria - Gestão de	507276078	EM
Figueira da Foz	501305580	Figueira parques - Estacio- namento	507276078	EM
Figueira de Castelo	505987449	Figueira Cultura e Tempos Livres	504766961	EM
Funchal	511217315	Sociohabitafunchal, EM de Habitação	511237880	EM
Funchal	511217315	EIMRAM - Investimentos e Serviços	511144121	EIM
Funchal	511217315	Sociohabitafunchal	511237880	EM
Fundão	506215695	Fundão Turismo	500645035	EM
Fundão	506215695	Viverfundão	507197895	EM
Gondomar	506848957	Gondomar Coração de Ouro	508252393	EM
Gouveia	506510476	D L C G - Desporto Lazer e Cultura	506510913	EM
Grândola	506823318	Infratroia - Infraestru- turas	505263963	EM
Guarda	501131140	Culturguarda	507210557	EM
Guarda	501131140	Guarda Cidade Desporto	504456261	EM
Guarda	501131140	Guarda	680018816	SMAS
Guimarães	505948605	Vimágua - Água e Saneamento	505993082	EIM
Guimarães	505948605	CASFIG - Coordena- ção das	504885855	EM
Guimarães	505948605	Vitrus Ambiente	509584888	EM SA
Guimarães	505948605	AVEPARK - Parque de Ciência e Tec.	506818934	SA
Horta	512073821	Urbhorta	512090084	EEM
Horta	512073821	HORTALUDUS - Gestão de	512076170	EM
Lagoa (R.A.A)	512074410	EML - Empresa Municipal Urban.	512090769	EM
Lagos	505170876	Fut*url*agos - Desenvolvimento	507684532	EEM
Lagos	505170876	Lagos-em-Forma - Gestão Desportiva	507725077	EEM
Lajes do Pico	512074143	Culturpico	512095841	EM
Lamego	506572218	Lamego Convida - Gestão de Equip.	507768060	EEM
Leiria	505181266	Leirisport - Despor- to, Lazer e Turismo	505183692	EM
Leiria	505181266	Leiria	680017550	SMAS
Lisboa	500051070	EPUL - Urbanização de Lisboa	500906475	EM
Lisboa	500051070	Lisboa Ocidental SRU Sociedade de	507023129	EM
Lisboa	500051070	Companhia Carris de Ferro de Lisboa	500595313	EM SA
Lisboa	500051070	EGEAC - Gestão de Equip. e Animação	503584215	EM SA
Lisboa	500051070	EMEL - Emp. Púb. Mun. de Mob.	503311332	EM SA
Lisboa	500051070	GEBALIS - Gestão Bairros Munic. de	503541567	EM SA
Loulé	502098139	Infralobo - Empresa de infra-estruturas	504041193	EM
Loulé	502098139	Inframoura - Empresa de Infra- -	504915266	EM
Loulé	502098139	Infraquinta - Emp. de Infra-Estruturas	503830704	EM SA
Loulé	502098139	Loulé Concelho Global	505493870	EM SA
Loures	501294996	Gesloures - Gestão de Equipamentos	502814063	EM
Loures	501294996	Loures Parque – EM de Estacionamento	505072947	EM
Loures	501294996	VALORSUL - Valorização e	509479600	SA
Loures	501294996	Loures	680009671	SM

Lousada	505279460	Lousada Seculo XXI – Activ. Desport. e	505840464	EM
Machico	511239440	Viver Machico	509867162	EEM
Madalena	512070946	Madalena Progresso	512095094	EEM
Madalena	512070946	Madalenagir	512099642	SA
Mafra	502177080	Mafratlântico - Vias Ro- doviárias	505216329	EM
Mafra	502177080	Pavimafra - Infra- -Estruturas e	505216329	EM
Mafra	502177080	Giatul - Gest. Infra-Estruturas em	506874915	EM SA
Maia	505387131	Maiambiente	505060868	EEM
Maia	505387131	Empresa Metropolitana de Estaciona-	504830783	EM
Maia	505387131	Espaço Municipal - Renov. Urb e Gest.	505462583	EM
Maia	505387131	TECMAIA - Parque de Ciência e	504569244	EM
Maia	505387131	EEA - Empresa de Engenharia e	506410803	SA
Maia	505387131	Maia	680015124	SME-
Maia	505387131	Electricidade, Água e Saneamento	680015124	SMEAS
Marinha Grande	505776758	TUMG - Transportes Ur- banos da	505849348	EM
Matosinhos	501305912	Matosinhoshabit - Hab. DE Matosinhos	504597221	EEM
Matosinhos	501305912	MS Matosinhos Sport	506197174	EM SA
Mealhada	506792382	EPVL - Escola Profi ssional da	504547313	Soc.
Melgaço	505592940	Cura Aquae - Termas de Melgaço	509688373	EM
Melgaço	505592940	Melsport - Melgaço Des- porto e Lazer	505922274	EM
Mértola	503279765	Merturis - Turismo	506888460	EEM
Miranda do Douro	506806898	Miranda Cultural e Rural	507174763	EM
Mirandela	506881784	AIN - Agro-Indús- trial do Nordeste	503193259	SA
Mirandela	506881784	Mirandela	680042423	SMA
Montijo	502834846	Montijo	680047930	SMAS
Moura	502174153	Herdade da Contenda	509455484	EM
Moura	502174153	Lógica - Sociedade Gesto- ra do Parque	508201306	EM
Nazaré	507012100	Nazaré	680017399	SM
Nordeste	512042659	HSN - Habitação Social do Concelho de	512090319	EM
Nordeste	512042659	Nordeste Activo	512088357	EM SA
Óbidos	506802698	Óbidos Patrimonium - ges- tão e	506916170	EEM
Óbidos	506802698	Óbidos Criativa	507566343	EM
Odivelas	504293125	Municipália - Gestão de Equipa- mentos	506219992	EM
Odivelas	504293125	Loures e Odivelas	680009671	SIMAR
Odivelas	504293125	Odivelas e Loures	680009671	SIMAR
Oeiras	500745943	Oeiras Viva - Gest. Equip. Sócio Cult. e	505351064	EEM
Oeiras	500745943	Lemo - Laboratório de Ensaios de	508424780	EIM
Oeiras	500745943	Parques Tejo - Par- queamentos de	504719670	EM
Oeiras	500745943	Oeiras e Amadora	680015019	SIMAS
Oeiras	500745943	HABITÁGUA - Servi- ços	503172022	Soc.
Olhão	506321894	Fesnima - Animação de Olhão	504667521	EEM
Olhão	506321894	SRU-Reabilitação Urbana de Olhão	501460888	EEM
Olhão	506321894	Ambiolhão - Ambiente de Olhão	509680780	EM
Olhão	506321894	Mercados de Olhão	504288865	EM
Oliveira de	506302970	Gedaz - Gestão de Equipa- mentos	508954703	EEM

Ourém	501280740	Ambiourem - Gestão de Espaços e	505765500	EEM
Ourém	501280740	SRU fatima - Sociedade de Reabilitação	507273885	EEM
Ourém	501280740	Verourem - Gestão de Equipamentos	505020408	EEM
Ourém	501280740	Ourém Viva - Gest. Eventos, Serviços e	505111691	EM SA
Ovar	501306269	Ovar Forma - Ensino e Formação	504599550	EM
Paços de Ferreira	502173297	Gespaços - Gestão de Equipamentos	505317982	EM
Paços de Ferreira	502173297	PFR Invest - Gestão Urbana	508278279	EM
Palmela	506187543	Palmela Desporto	504706675	EM
Paredes	506656128	Amiparedes - Agência Mu- nicipal de	508799821	EM SA
Penafi el	501073663	Penafi el Activa	506196917	EEM
Penafiel	501073663	Penafiel Activa	506196917	EEM
Penafiel	501073663	Penafiel Verde - Entidade Empresarial	507700651	EEM
Peniche	506812820	Peniche	680019600	SMAS
Pinhel	506787249	Falcão Cultura Turismo e Tempos	507742834	EM
Pombal	506334562	Pombal Prof - Sociedade de Educação e	504609696	Soc.
Ponta Delgada	512012814	Azores Parque	512081727	EM SA
Ponta Delgada	512012814	Cidade em Acção	512088845	EM SA
Ponta Delgada	512012814	Coliseu Micaelense - Soc. Prom.	512059420	EM SA
Ponta Delgada	512012814	Ponta Delgada	672001721	SMAS
Portalegre	501143718	Agua e Transporte de Portalegre	680031065	SMAT
Portimão	505309939	EMARP - Águas e Resíduos	505322730	EM
Portimão	505309939	Portimão Renovada SRU - Sociedade de	501460888	EM
Portimão	505309939	Portimão Turis	508666236	EM
Portimão	505309939	Portimão Urbis SGRU – Gestão	505574233	EM SA
Porto	501306099	CMPEA - Águas do Porto	507718666	EM
Porto	501306099	CMPH - Domus Social - Emp. Habit. e	505037700	EM
Porto	501306099	CMPL - Porto Lazer	507718640	EM
Porto	501306099	DOMUSSOCIAL – Emp. Habitação e	505037700	EM
Porto	501306099	Empresa Municipal de Ambiente do	514280956	EM
Porto	501306099	Gestão de Obras Públicas	505037238	EM
Porto	501306099	Porto Lazer	507718640	EM
Porto	501306099	APOR - Agência para a Modernização	508184509	SA
Porto	501306099	Portovivo, SRU - Sociedade de Rea-	506866432	SA
Porto	501306099	PRIMUS - Promoção e	504558161	SA
Póvoa de Lanhoso	506632920	EPAVE - Escola Profi ssional do Alto	504596608	Soc.
Póvoa de Varzim	506741400	Varzim Lazer	504841700	EEM
Povoação	512065047	Espaço Povoação - Ac- tividades	512085668	EM
Povoação	512065047	POVOAINVEST - Empresa Municipal	512084947	EM
Praia da Vitória	512044023	Praia em Movimento	512099472	EM
Praia da Vitória	512044023	TERAMB	509620515	EM
Praia da Vitória	512044023	Praia Ambiente	512097780	EM SA
Proença-a-Nova	505377802	Proençatur - Empresa de Turismo	505396114	EM
Ribeira Grande	512013241	Musami - Operações Mu- nicipais do	512096481	EIM
Ribeira Grande	512013241	Ribeira Grande Mais - Ha- bitação	512086338	EM
Rio Maior	505656000	Desmor - Gestão Des- portiva	504748114	EM

Rio Maior	505656000	EPRM - Escola Profi ss. de Rio Maior	504617656	Soc.
Sabugal	506811662	Sabugal - Gestão de Espaços Culturais,	502726245	EM
Santa Comba Dão	506637441	Profi academus - Escola Profi ssional	504609718	Soc.
Santa Maria da	501157280	Feira Viva Cultura e Desporto	505120151	EEM
Santa Maria da	501157280	PEC-tSM - Parque Empre- sarial da	900220538	EM
Santa Maria da	501157280	Indaqua Feira - Indústria e Gestão de	504520890	SA
Santa Maria da	501157280	Sociedade de Turismo de Santa Maria da	508905435	SA
Santarém	505941350	Cul.tur - Empresa Munici- pal de Cultura	509477755	EEM
Santarém	505941350	Lt - Sociedade de Reabili- tação Urbana	509226426	EM
Santarém	505941350	Scalabisport - Gestão de Equipa- mentos	506159540	EM
Santarém	505941350	STR-URBHIS - Sociedade de Gestão	509472087	EM SA
Santarém	505941350	Viver Santarém	506159540	EM SA
Santo Tirso	501306870	Electricidade, Água e San. de Santo	680019391	SMEAS
São João da	506538575	Habitar S. João - Entidade Empresarial	506546365	EEM
São João da	506538575	Perm - Parque Empresa- rial de	509042201	EIM
São João da	506538575	Águas de João	508326567	EM SA
São Pedro do Sul	506785815	Termalistur - Termas	506817997	EEM
São Roque do Pico	512074771	Cais Invest	512097666	EEM
São Vicente	511240112	Naturnorte – Gest. Equip. Colect. e	511086040	EM SA
Seia	506676170	EMCR - Cultura e Recreio	505703262	EM
Sernancelhe	506852032	ESPROSER - Escola Pro- fi ssional	504676326	SA
Setúbal	501294104	CDR - Cooperação e Desenvolvimento	600076849	SA
Sever do Vouga	502704977	Vougapark - Parque Tec- nológico e de	507760476	EM
Sintra	500051062	Educa - Gestão e Manu- tenção de	504845535	EEM
Sintra	500051062	Sintra Quorum - Gestão de	504605062	EEM
Sintra	500051062	EMES - Estacionamento	504610163	EM
Sintra	500051062	HPEM - Higiene Pública	505031280	EM
Sintra	500051062	Sintra	680000054	SMAS
Tavira	501067191	Tavira Verde - Ambiente	507236335	EM
Tomar	506738914	Tomar	680039457	SMAS
Tomar	506738914	Ensino Profissional de Tomar	504699326	Soc.
Torres Novas	506608972	TMTN - Teatro Municipal	507695259	EM
Torres Vedras	502173653	Promotorres - Prom. Eventos e Gestão	503941565	EM
Torres Vedras	502173653	Torres Vedras	680015973	SMAS
Trofa	504296434	Trofa Park - Reabilit. Urb., Desenv.	506788830	EEM
Trofa	504296434	Trofaguas - Serviços Ambientais	506236838	EM
Valença	506728897	Interminho – Soc. Gestora de Parques	504923242	EM
Valongo	501138960	Vallis Habita - Gestão de Emp. Hab.	505265800	EM
Velas	512075506	Velasfuturo - Gestão de Equipamentos	512098239	EM
Viana do Alentejo	506151174	Viana do Alentejo	509133843	SMSB
Viana do Castelo	506037258	Viana Castelo	680012907	SMSB
Vieira do Minho	506659682	Vieira Cultura e Turismo	500792933	EM
Vila da Praia da	512044023	Praia Ambiente	512097780	EM
Vila da Praia da	512044023	Praia em Movimento	512099472	EM
Vila do Porto	512063770	SDMSA - Sociedade Desenvolvimento	509374255	EEM

Vila Franca de Xira	506614913	Vila Franca de Xira	680021892	SMAS
Vila Nova de Foz	506829197	Fozcoactiva - Gestão Equipamentos	506383407	EM
Vila Nova de Gaia	505335018	Águas de Gaia	504763202	EEM
Vila Nova de Gaia	505335018	Cidadegaia - Sociedade de Reabilitação	507244419	EEM
Vila Nova de Gaia	505335018	Gaianima	505336405	EEM
Vila Nova de Gaia	505335018	Gaiasocial - Habitação	504430823	EEM
Vila Nova de Gaia	505335018	Gaiaurb - Gestão Urbanística e da	506064433	EEM
Vila Nova de Gaia	505335018	Parque Biológico de Gaia	504888773	EEM
Vila Pouca de	506810267	Vitaguiar - Apoio ao Desenvolvimento	507381769	EM
Vila Real	506359670	CULTURVAL - Gestão de Equip	506644782	EM
Vila Real	506359670	EMARVR - Água e Residuos de Vila	506516725	EM
Vila Real	506359670	MERVAL - EM de Gestão de Merc e	505324024	EM
Vila Real	506359670	VRS - Social, Habitação e Transportes	506376745	EM
Vila Real de Santo	506833224	VRSA - Sociedade de Ges- tão Urbana	508160570	EM SA
Vila Verde	506641376	Proviver	507914465	EM
Vila Verde	506641376	Escola Profi ssional Amar Terra Verde	504595067	Soc.
Vimioso	506627888	Caça e Turismo de Vimioso	506976246	Soc.
Vinhais	501156003	ProRuris – Desenvolvimento Rural	507643720	EEM
Vinhais	501156003	Turimontesinho – EM de Promoção	507647130	EEM
Viseu	506697320	Habisolvis - Habitação Social	506804186	EM
Viseu	506697320	Viseunovo - SRU - So- ciedade de	507406672	SA
Viseu	506697320	Viseu	680020063	SMAS
Vizela	505985217	Vimágua - Água e Saneamento	505993082	EIM SA